PACLIC 36 (2022)

Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation

20-22 October, 2022
De La Salle University
Manila, Philippines

ISSN 2619-7782
Published by De La Salle University
2401 Taft Ave, Malate, Manila
1004 Metro Manila

Acknowledgement
PACLIC 36 (2022) is jointly hosted by the Department of English and Applied Linguistics (DEAL), Br. Andrew Gonzalez FSC College of Education (BAGCED), De La Salle University, Manila, Philippines, the College of Computing and Information Technologies (CCIT), National University, Manlia, Philippines, and the Lasallian Institute for Development and Education Research (LIDER), De La Salle University, Manila, Philippines.

Foreword

The 36th Pacific Asia Conference on Language, Information, and Computation, PACLIC 36 (2022) is organized by the Department of English and Applied Linguistics, Br. Andrew Gonzalez FSC College of Education, De La Salle University, the National University (Philippines), and the Lasallian Institute for Development and Educational Research (LIDER), De La Salle University. This edition of the PACLIC series of conferences, as its long tradition, continues to emphasize the synergy of theoretical analysis and processing of natural language, aiming to strengthen the communication between researchers working in different fields of language study in the Asia-Pacific region as well as around the world.

Like its predecessors in 2020 and 2021, the 36th PACLIC Conference was held fully online due to the COVID-19 pandemic's continued restriction on mobility in the Philippines. We fervently hope that the local and global situation on COVID eases in the future, such that we could resume a face-to-face format of the conference. We received 164 submissions out of 25 countries in total, tallying a 62.19 percent acceptance rate. In the 102 papers accepted, 84 of which were for oral presentations and 18 for poster presentations. In addition to excellent oral and poster presentations, the conference also highlighted three plenary talks and three invited talks delivered by notable scholars in the field. We thank Kathleen Ahrens, Hanjung Lee, Stefanie Shamila Pillai, Wilkinson Daniel Wong Gonzales, and Nguyen Thi Minh Huyen.

We would also like to thank the PACLIC Organizing and Steering Committee for being the backbone of this conference, especially in light of logistical difficulties and unique challenges of the conference's online setup. We express our sincerest gratitudes to the Local Organizing Committee, Reviewers, and the Secretariat for ensuring that the conference's preparation and execution were done efficiently and smoothly. Lastly, we would like to thank Department of English and Applied Linguistics, the Br. Andrew Gonzalez FSC College of Education and the Office of the President and the Office of the Vice President Research and Innovation of the De La Salle University for all the support to this endeavor.

Shirley N. Dita
Jong-Bok Kim
Rachel Roxas
36th PACLIC Program Committee Chairs
(on behalf of the Organizing Committee)

Organizers

PACLIC Steering Committee Standing Members:

Chu-Ren Huang, The Hong Kong Polytechnic University, Hong Kong
Jong-Bok Kim, Kyung Hee University, Seoul
Ryo Otoguro, Waseda University, Tokyo
Rachel Edita O. Roxas, National University, Manila
Maosong Sun, Tsinghua University, Beijing
Benjamin T'sou, City University of Hong Kong, Hong Kong
Min Zhang, Soochow University, Suzhou

Local Organizing Committee

Conference Chair: Shirley N. Dita
Conference Co-Chairs: Arlene O. Trillanes
Rochelle Irene G. Lucas

Program Committee

Chairs: Jong-Bok Kim
Rachel Editha Roxas
Shirley N. Dita
Co-Chairs: Rochelle Irene Lucas
Leah Gustilo
Raymund Vitorio
Members: Sharon Joy Bulalang
Grace Cortez
Aileen Bautista

Publicity Committee

Chair: Mideth B. Abisado
Members: Angelique D. Lacasandile
Ryan Richard H. Guadaña

Registration Committee

Chair: Marilou Santos
Member: Ma. Angelica Gumangan

Logistics Committee

Chair: Frederick Talaue
Members: Philip Rentillo
Aileen Bautista

Communications Committee

Chair: Jean Purpura
Member: Eden Grace Canopio

Secretariat

Lovely Liz De Luna
Ma. Nilda Perilla
Ronnie Sabado

Program Sub-committee/Reviewers:

Muhammad Afzaal
Ericson Alieto
Alice Mae Arbon
Aireen Arnuco
Chirbet Ayunon
Judith Azcarraga
Nguyen Bach
Aileen Bautista
Alejandro Bernardo
Annie Mae Berowa
Sharon Bulalang
Mary Joy Canon
Marvin Casalan
Jasper Kyle Catapang
Angie Ceniza
Maria Art Clariño
Romina Grace Cortez
Kristine de Leon
Aprillette Devanadera
Nimfa Dimaculangan
Chenee Dino-Aparicio
Shirley Dita
Jennibelle Ella
Ma. Regina Estuar
Ramsey Ferrer
Ana Cristina Fortes
Cecilia Genuino
Ryn Jean Fe Gonzales
Wilkinson Gonzales
Yanhui Gu
Nicanor Guinto
Munpyo Hong
Joseph Imperial
Rowland Imperial
Jeffrey ingosan
Jong-Bok Kim
Valia Kordoni
Huei-Ling Lai
Nguyen Le Minh
Phuong Le-Hong
Yong-Hun Lee
Xiaoqian Li
Joanna Marie Lim
Rochelle Irene Lucas
Romualdo Mabuan
Alvin Malicdem
Erlyn Manguilimotan
Wenwei Mao
Vladimir Mariano
Dalos Miguel
Eusebio Jr. Mique
Alen Mateo Muñoz
Naonori Nagaya
Thi Minh Nguyen
Nathaniel Oco
Ethel Ong
Chautamanee Onsuwan
Jong C. Park
Rafael Michael Paz
Charmaine Ponay
Ariel Robert Ponce
Nattama Pongpairoj
Jeanne Purpura
Edward Jay Quinto
Rodolfo Jr. Rafa
Reginald Neil Recario
Philip Rentillo
Rachel Roxas
Byong-Rae Ryu
Ria Sagum
Julie Ann Salido
Sheena Sapuay
Shu-Ing Shyu
Melanie Siegel
Pornsiri Shinghapreecha
Sanghoun Song
Elineth Suarez
Christian Sy
Jennifer Tan-De Ramos
Michael Tanangkingsing
Zhivang Teng
Gina Ugalingan
Paolo Valdez
Raymund Vitorio
Tak-Sum Wong
Hongzhi Xu
Cheng-Zen Yang
Aiden Yeh
Satoru Yokoyama
Liang-Chih Yu

Invited Talks

Keynote Speaker

Breaking through glass ceilings or opening glass doors? Addressing the challenges of equality and inclusion via metaphors

Kathleen Ahrens (The Hong Kong Polytechnic University)

Women, but not men, are faced with metaphorical glass ceilings, and in some cases, glass cliffs. These novel constructed metaphors reflect the difficulties women have in advancing in their career, particularly in terms of advancement in leadership roles in politics and in the professions. However, what is less clear is if there are gendered associations for source domains associated with conventionalized conceptual metaphors. This talk will examine this issue and present recent advances in conceptual metaphor theory that use ontologies and collocational patterns to verify source domains (Ahrens &amp; Jiang 2020) as a prior step to examining gendered associations. Drawing on data from business and politics, I will also explain how women use these conventionalized metaphors to position themselves as leaders. I will close my talk with a discussion of how we, as scholars, can influence the creation and interpretation of metaphor use so as to engender positive social change.

Plenary Speakers

Cue reliabillity and motivation in grammar and language use: A new look at differential subject marking

Hanjung Lee (Sung Kyun Kwan University, South Korea)

In many languages, arguments such as subjects and objects enjoy substantial freedom in terms of the form in which they are realized: a noun or pronoun with or without a following or preceding functional particle, a bare noun, a null-form argument, etc. This talk is concerned with caseless subjects in Korean, that is, those subjects that occur without functional particles signaling case or discourse function. An interesting and challenging problem for theoretical approaches to case is that caseless and case-marked subjects are in systematic contrast as to their interpretation. This makes Korean a Differential Subject Marking (DSM) language in Aissen's (2003) terms, wherein some subjects are marked with formal particles while others are not, depending on the semantic and pragmatic features of the subject.

This talk will focus on a hitherto unexplained property of caseless-subject clauses triggering a direct perception interpretation. I will first present evidence from conversation data demonstrating that caseless subjects predominantly occur in clause types that have an agent directly identifiable in the here and now, whereas nominative-marked subjects are most productively used in clause types wherein the identification of an agent cannot be grounded in the here and now. Based on this evidence, I will propose a new account of DSM in terms of cue reliability (Rosch & Mervis 1975; Levshina 2021), arguing that the association of caseless subjects with seemingly unrelated features such as direct perception in the here and now, agentivity, definiteness, tense deficiency and a simple, thetic interpretation (the most preferred information structure associated with caseless-subject clauses) follows from an economical use of formal particles motivated by the reliability of cues for identifying or predicting the grammatical, referential, or discourse status of an argument NP or a clause containing it.

When such cues are weak, sentences are likely to convey information that is less predictable and redundant; speakers prefer to mark the subject by overt particles in such sentences, as using case markers would lead to more uniform information density than leaving them unmarked.

These preliminary results support efficiency-based accounts of patterns of grammar, and underscore the importance of communicative efficiency in explaining and motivating grammatical rules and language use (Hawkins 2004, Haspelmath 2008, Jaeger 2010, Lee 2010, 2016, 2021, Lestrade & de Hoop 2016, Levshina 2021).

Building the next generation WordNet for Filipino with sense embeddings and network science

Briane Paul V. Samson (De La Salle University, Manila, Philippines)

The low-resource and highly morphological setting of Philippine languages is a challenge in developing a word representation or a language model. The current linguistic resources lack in rich semantic data that is crucial in most NLP tasks, and the fast-paced evolution and adaptation of Philippine languages make things even more difficult in creating a well defined language model. Thus, building a Filipino WordNet is crucial in advancing the landscape of Filipino NLP. With the vast amount of data in many digital platforms that can represent different domains and varieties of words through time including changes in its semantic and syntactic forms, we aim to create word representations of the Filipino language that is temporal and context aware and store them in the expanded Filipino WordNet. Given that languages continue to evolve and adapt, we also investigate the diachronic emergence and semantic shifts of word senses across different contexts and media, especially for the low-resource Filipino and Philippine English languages.

Invited Lectures

BibePortMal: A mobile app dictionary for Melaka Portuguese

Stefanie Shamila Pillai (University of Malaya, Malaysia)

Melaka Portuguese (also known as Papiá Cristang) has its roots in the arrival of the Portuguese in Melaka in the16th Century. The language is still spoken by the descendants of unions between the Portuguese and locals especially in Melaka, but the dwindling number of fluent speakers and the lack of intergenerational transmission has led to this language being classified as one of the many endangered languages in Malaysia (Pillai, Soh, & Kajita, 2014). Several community engagement projects have been undertaken to translate research into efforts to revitalise Melaka Portuguese (Pillai, Phillip, & Soh, 2017). Among the efforts to encourage the use of the language and to assist the teaching of the language by a community member, a Melaka Portuguese-English dictionary in the form of a mobile application, BibePortMal, was developed. In this session I will talk about the rationale and the process of developing the application with community representatives. Some of the challenges in its development will also be discussed along with plans to improve the application.

Harnessing the power of social media in linguistic analysis: A diachronic and sociolinguistic study of Philippine English(es) using the Twitter Corpus of Philippine Englishes

Wilkinson Daniel Wong Gonzales (The Chinese University of Hong Kong)

Research on contemporary Philippine English remain relatively scarce and inadequate in comparison to research on other varieties such as American English and Singapore English, partially due to the lack of large-scale, organized, publicly available data sets that allow comprehensive and in-depth investigations of the variety. Responding to this demand, I introduce the Twitter Corpus of Philippine Englishes (TCOPE) – a 135-million-word corpus created from roughly 27 million tweets sampled from 29 major cities in the Philippine archipelago. In the first part of the talk, I provide an overview: I discuss the considerations that went into TCOPE's design, the compilation procedure, the format, and how interested individuals can access the corpus. Then, I illustrate the utility of the corpus by showcasing how it can be used to insightfully examine the linguistic features of Philippine English as well as the relationship between these features and socio-temporal factors (e.g., ethno-geographic region, time, age, sex), focusing on four documented Philippine English features: (1) the use of irregular past tense morpheme -t, (2) double comparatives, (3) subjunctive were in subordinate counterfactual clauses, and (4) the phrasal verb base from. My initial explorations confirm patterns observed in previous research but go further to show the multifaceted and dynamic nature of Philippine English, providing empirical support for the theory that Philippine English is at the final stage of Schneider's dynamic model. Because of its large size, sampling distribution, and its availability in different corpus formats, TCOPE can be used to investigate features in 'general' contemporary Philippine English as well as different types of variation, particularly diachronic and ethno-geographic variation – a feat that might not be possible with other Philippine English corpora. In combination with other existing corpora, TCOPE has the potential to broaden horizons in the diachronic and sociolinguistic study of Philippine English(es).

Towards universal syntactic-semantic resources for Vietnamese

Nguyen Thi Minh Huyen (VNU University of Science, Hanoi, Vietnam)

Despite the fact that Vietnamese is spoken by around 100 million persons all over the world, it remains a low resource language in terms of gold datasets for natural language processing (NLP). Although many NLP applications have been developed in recent years thanks to the emergence of deep learning and embedding methods, it is important to build sustainable linguistic resources, using sophisticated linguistic annotation frameworks for the Vietnamese language, in harmony with universal frameworks. In this talk, I will present our work on the construction of Vietnamese syntactic-semantic resources, including a VerbNet-based lexicon and annotated corpora for syntactic and semantic parsing.

Table of Contents

Integrating Label Attention into CRF-based Vietnamese Constituency Parser

Duy Vu-Tran, Phu-Thinh Pham, Duc Do, An-Vinh Luong and **Dien Dinh**
University of Science, Ho Chi Minh city, Vietnam
Vietnam National University, Ho Chi Minh city, Vietnam
{vtduy18, phpthinh18}@apcs.fitus.edu.vn
{dotrananhduc, anvinhluong}@gmail.com
ddien@fit.hcmus.edu.vn

Abstract

Attention mechanisms and linear-chain conditional random field (CRF) have been applied to constituency parsing, and the achieved results are phenomenal. While self-attention and label attention layers (LAL) have been proven to be state-of-the-arts in English constituency parsing for their improvement in the encoding phase, the CRF two-stage technique shows its effectiveness in lowering computational cost. Attention-based architectures allow a word (self-attention) or a label (label-attention) to include its own viewpoint into extracted information. Our system is an extension of the current CRF-based model with additional attention-based methods to improve the quality of the encoding phase. Another crucial factor in our encoder is BERT as the pre-trained model has gained recognition in various natural language processing (NLP) tasks. Taking the advantage of different methods, we implement a model that combines label attention, contextualized encoding, and conditional random field. Furthermore, we adopt the biaffine attention, which is mainly used in the dependency parsing task, in our scoring layer. The architecture performs greatly on the Vietnamese treebank as it gives an over-85 F1-score on the test set and an over-82 F1-score on the dev set. On a larger scale, our idea of integration could be utilized in other language models.

1 Introduction

In the modern era, syntactical parsing has gained remarkable results due to the rise of deep learning and neural networks (Mrini et al., 2020; Zhang et al., 2020; Zhou and Zhao, 2019; Stern et al., 2017; Wang and Tu, 2020; Yang and Deng, 2020). Especially, constant improvements on English and Chinese constituency parsing come from proposals of machine learning techniques (Mrini et al., 2020; Zhang et al., 2020; Zhou and Zhao, 2019; Stern et al., 2017), which encourages us to apply such models to the same task of the Vietnamese language. Since most of the existing constituency parsers are encoder-decoder architectures, the main approach to improving the performance is to upgrade either encoder or decoder or both.

Encoders handle inputs and extract their significance in vector forms so that the model can easily understand them. Particularly, the inputs of a constituency parser are sentences, and the encoders try to learn the information of each word or span of the sentences. Recurrent neural networks (RNN) and Long short-term memory networks (LSTM) are the main tools to extract such features from data for their ability to learn the contextual characteristics, and Zhang et al. (2020); Stern et al. (2017); Gaddy et al. (2018) have benefited from these mechanisms. PHAN et al. (2019), one of the pioneers in Vietnamese constituency parsing, used BiLSTM in their encoder. Despite RNNs and LSTMs' ability to capture contextual information, they are surpassed by the works of self-attention.

In 2017, Vaswani et al. (2017) presented the model of Transformer and the self-attention mechanism, which opened a new chapter for natural language processing (NLP). Self-attention layers are capable to understand the global context of a given input and additionally, an attention-weighted view of the input's words to itself. With self-attention, (Kitaev and Klein, 2018) improved the works of (Stern et al., 2017; Gaddy et al., 2018). Later on, Tran et al. (2020) adopted the model for Vietnamese constituency parsing successfully, which inspires us to combine the architecture with our encoder.

After the emergence of Transformer, BERT (Devlin et al., 2019), which is a Transformer-based model and pre-trained on a large corpus of a target language, was proposed. PHAN et al. (2019); Tran et al. (2020) included PhoBERT (Nguyen and Nguyen, 2020) in their models as PhoBERT is specifically trained on Vietnamese, and impressive results are achieved. Besides that, PhoBERT also performed astonishingly in other Vietnamese NLP tasks (Nguyen and Nguyen, 2021) which we believe to enhance our encoder with the pre-trained knowledge of Vietnamese.

In 2020, Mrini et al. (2020) further developed the attention mechanism which resulted in Label Attention Layer (LAL). While self-attention (Vaswani et al., 2017) refers to input's views to itself, LAL offers the view of a label to the given sentence. Self-attention, BERT, and LAL are composed to better the encoder. The impact of LAL is further discussed in the Experiment section 3.

Besides dealing with the encoder, we adopt the conditional random field (CRF) concepts and two-stage decoding from Zhang et al. (2020) as they have proposed an efficient inside algorithm and proven the effectiveness of the methods. We also consider the biaffine attention (Dozat and Manning, 2017) for the scorer which is inspired by the task dependency parsing.

In this paper, we make a combination of current mechanisms into one single model: self-attention layers, label attention layers, PhoBERT for encoder; Biaffine Attention for scoring; two-stage CRF method for the decoder. We conduct examine our model on Vietnamese treebank (Nguyen et al., 2009) and achieve the results of over 85 on the test set and over 82 on the dev set (all results are given in F1-score). Section 2 re-describe our parser in the order of encoder, scorer, decoder, loss function. Section 3 presents the experiments' settings and a comparison between our model and other methods, and finally, we give our conclusion in section 4.

2 Model Architecture

2.1 Overview

Our parser includes 3 main components: encoder, scorer, decoder. Before being processed by the parser, the input sentence is transformed into the desired representations for the encoder. The encoder is a combination of the Attention mechanism and BERT, both of which extract essential information of the input sequence. On the Attention branch, the word-level weighted views for the sentence are extracted by k self-attention layers (Vaswani et al., 2017), and the following d_{lal}-head label attention(Mrini et al., 2020) is responsible for enhancing these outcomes with label-level weighted views. Additionally, the BERT(Devlin et al., 2019) branch handles the input independently and provides the features that are learned from the pre-training process. The results of two branches are aggregated into one sentence representation, which goes through the Biaffine scorer (Dozat and Manning, 2017) to compute the span score matrix, and the label score tensor. For the decoder, we apply the theory of Conditional Random Field (CRF) (Zhang et al., 2020) and lower the computational cost with the two-phase strategy. The model is referred to figure 1.

2.2 Encoder

Our encoder includes 2 parts: the Attention part and the BERT part. A visualization of the Attention part is given in figure 2. The idea of the encoder is based on the hypothesis that a constituency tree depends on the set of constituency labels and the context of the given text. To realize the assumption, we use:

- k self-attention layers to extract the information of how each word of a sequence 'see' the sequence's context itself.
- d_{lal}-head label-attention layer on the top of the self-attention layers to retrieve the viewpoint of the constituency labels to the input sentence.
- BERT on the other branch to improve the encoder with pretrained contextual features.

Token representations For the Attention part, we concatenate the content and the position embedding following (Vaswani et al., 2017), where the content is represented by the part-of-speech (POS) embedding [1]. We choose POS embedding instead of the word embedding because we observe that the POS one gives higher performance (which we will discuss in the Experiment section (3)) and because the information of words are extracted by the BERT layer instead.
For the BERT part, we tokenize the sequence into subword representations which are widely used for BERT encoders.

Self-attention Mechanism We follow the encoder implementation of (Mrini et al., 2020) in which the token representations are put through several self-attention layers (Vaswani et al., 2017) before being processed by the LAL.

Self-attention consists of a number of consecutive layers, each of which has identical operations.

[1] The POS tags are in XPOS forms.

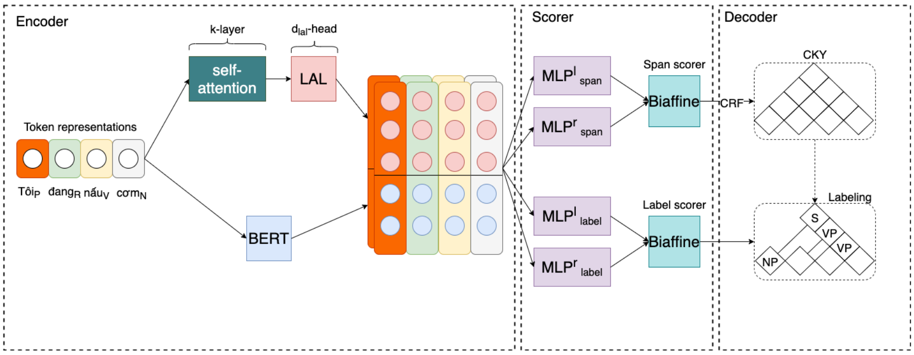


Figure 1: Model architecture. The main flow of the model for the sentence "Tôi đang nấu cơm". The sentence is encoded by Label Attention Layer (Mrini et al., 2020) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Afterwards, it is moved through MLP to extract features before being scored by the Biaffine (Dozat and Manning, 2017) mechanism and Conditional Random Field (CRF) (Zhang et al., 2020). The score is decoded using Cocke–Kasami–Younger (CKY) algorithm.

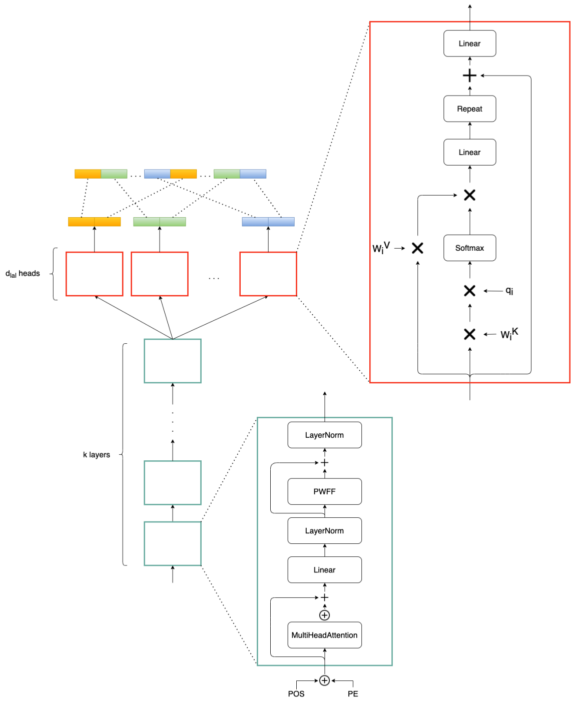


Figure 2: The encoder of attention mechanism includes 2 parts. Firstly, The token representations go through k self-attention layers to extract the attentive-weighted information. Subsequently, the data are fed into d_{lal}-head label attention. The output of each head is divided into 2 halves which act as forward and backward features of the word. (The $\times$ operator refers to the matrix multiplication operation, and the $\oplus$ implies the concatenation. POS and PE stand for part-of-speech embedding and positional embedding).

In a layer k of the self-attention, for a single head j in the self-attention layers, given the input matrix $X^{(k)}$:

$$c'_j = softmax(\frac{(W_j'^Q X^{(k)})(W_j'^K X^{(k)})}{\sqrt{d'}})W_j'^V X^{(k)} \tag{1}$$

where $W_j'^Q, W_j'^K, W_j'^V$ are the learnt matrices of the head j. We used $'$ to distinguish notations between the label attention and the original self-attention. Each attention head j learns from the input individually as they have their own trainable parameters. The chosen number of heads is 8 following (Vaswani et al., 2017).

To aggregate multiple heads of attention, we concatenate the outputs together:

$$output' = Linear(c'_1 \oplus c'_2 \oplus ... \oplus c'_8) \tag{2}$$

where the $Linear$ operation is used to transform the output back to the input X 's dimension. $Multihead$ is followed by a residual connection and a Layer Normalization (Ba et al., 2016) which results in the layer k's output.

$$\begin{aligned} L^{(k)} &\equiv Layer^{(k)}(X^{(k)}) \\ &= LayerNorm(X^{(k)} + output') \end{aligned} \tag{3}$$

Before being fed to the next layer $k+1$, the output goes through a position-wise feed-forward which has the same form as in (Vaswani et al., 2017) and a residual-LayerNorm again:

$$X^{(k+1)} = LayerNorm(L^{(k)} + PWFF(L^{(k)})) \tag{4}$$

where PWFF is the position-wise feed-forward function.

$$PWFF(X) = Linear(ReLU(Linear(X))) \tag{5}$$

where ReLU refers to the Rectified Linear Unit function. The input and output of position-wise feed-forward have the same dimensions (Kitaev and Klein, 2018).

Label Attention Layer For English constituency parsing, the authors of (Mrini et al., 2020) have proposes the Label Attention Layer(LAL), and it has produced a state-of-the-art result. As Label Attention allows the labels to give their attention views on a given sentence, we apply the mechanism to our model as a part of encoder. A comparison between models with and without LAL is made and reported in the Experiment section (3).
The LAL takes a matrix X as input which consists of embedded vectors of words of a given sentence. The attention weight of each head j is represented by an attention-weight vector a_j which is calculated from the query vector q_j.

$$a_j = softmax(\frac{q_j(W_j^K X)}{\sqrt{d}}) \tag{6}$$

where W_j^k is a learnt matrix of head j, and it is used to learn the key factors of X. d is the length of the query vector. Afterwards, the context vectors c_j are computed (Label Attention Layer uses vectors instead of matrices).

$$c_j = a_j(W_j^V X) \tag{7}$$

where W_j^V is the learnt value matrix of head j.

The context vectors are projected to the dimension X's vectors before being repeatedly added to X.

$$c_j = Linear(c_j) \tag{8}$$

$$output_j = \begin{bmatrix} c_j \\ c_j \\ \cdots \\ c_j \end{bmatrix} + X \tag{9}$$

The $output_j$ is a matrix representing the label j view of the sequence X, and it is projected to a smaller dimension d_{lal} to prevent overfitting as well as to optimize computational effectiveness. The outcomes of all attention heads j (for $j \in [1; d_{label}]$), given that d_label is the number of heads), are concatenated into the final output of LAL.

$$output_j = Linear(output_j) \tag{10}$$

$$\mathbf{L} = output_1 \oplus output_2 \oplus ... \oplus output_{d_{label}} \tag{11}$$

$\mathbf{L} \in \mathbb{R}^{n\times(d_{label}*d_{lal})}$ where n is the length of the given sequence, and $\mathbf{L}_i$ represents the vector of the given sequence's i^{th} word.

Although the paper (Mrini et al., 2020) states that we can choose as many attention heads as the labels, they also show that the relation between the number of attention heads and one of the labels is not specifically one-to-one. This number is a hyper-parameter that should be chosen via experiments which we present in the Experiment section(3). The LAL is put on top of the self-attention module to provide the labels' weighted views for the words' attentions. In other words, the self-attention layers are not replaced by the LAL but instead, are enhanced by it.

Inside the self-attention and label attention layers, we apply the partition for content embedding and positional embedding as it is shown in (Kitaev and Klein, 2018) to gain better performance.

BERT Fine-tuning BERT (Devlin et al., 2019) is a context-aware embedding encoder, and it is pretrained on a large corpus of the target language. We decide to fine-tune the encoder further which leads the pretrained model to fit the treebank. The authors of (Nguyen and Nguyen, 2020) have introduced a RoBERTA-based (Liu et al., 2019) model: Phobert, and the model is pretrained on a large Vietnamese corpus, which makes it more suitable for the task. PhoBERT (Nguyen and Nguyen, 2020) has shown impressive improvement in many Vietnamese tasks such as dependency parsing, POS tagging, named entity recognition (Nguyen and Nguyen, 2021). Given a sequence of subwords $\{sw_1, sw_2, ..., sw_n\}$, the output of BERT is denoted as:

$$\mathbf{B}_i = BERT(sw_i) \tag{12}$$

While the LAL provides the viewpoint of the labels to a given delexical sequence, BERT supports the contextual attention views of each word to the input sentence.

Input Representation Following (Stern et al., 2017), we combine the forward and backward representations of the output. We split the composition of BERT and LAL in two halves and treat them as the forward and backward representations. For a n-long sequence s, the process is shown in equations

13, 14, 15.

$$\mathbf{for}_i = \mathbf{B}_i[0:n/2] \oplus \mathbf{L}_i[0:n/2] \tag{13}$$

$$\mathbf{back}_i = \mathbf{B}_i[n/2+1:n] \oplus \mathbf{L}_i[n/2+1:n] \tag{14}$$

$$\mathbf{E}_i = \mathbf{for}_i \oplus \mathbf{back}_{i+1} \tag{15}$$

2.3 Scorer

Inspired by the Biaffine Attention (Dozat and Manning, 2017) used for the task of dependency parsing, we apply the mechanism to our scoring architecture. The authors of (Zhang et al., 2020) made a comparison between the Biaffine scoring method with the previous one (Stern et al., 2017), and the Biaffine one gave a consistently higher performance.

Feature Extraction We extract the information of the left and right boundaries using two separate MLP layers as each word w_i acts as either the left boundary (the span endpoint is to the left of w_i) or the right boundary (the span endpoint is to the right of w_i) of a span. Another two MLP layers are used for the labelling task.

$$\mathbf{span}^l_i, \mathbf{span}^r_i = MLP^l_{span}(\mathbf{E}_i), MLP^r_{span}(\mathbf{E}_i) \tag{16}$$

$$\mathbf{label}^l_i, \mathbf{label}^r_i = MLP^l_{label}(\mathbf{E}_i), MLP^r_{label}(\mathbf{E}_i) \tag{17}$$

Biaffine Scorer The Biaffine takes the outputs from the feature extraction process as it inputs. Particularly for a span (i, j), the Biaffine uses the left boundary features of the i^{th} word and the right one of the j^{th} to score the span. Similarly, we gain the label scores of any span (i, j) and label $l \in \ell$.

$$s(i,j) = \begin{bmatrix} \mathbf{span}^l_i \\ 1 \end{bmatrix}^T \mathbf{D} \left[\mathbf{span}^r_j\right] \tag{18}$$

where $\mathbf{D}$ is a trainable parameter, and $\mathbf{D} \in \mathbb{R}^{d \times d}$. d is the output dimension of $MLP^{l/r}_{span}$.

$$s(i,j,l) = \begin{bmatrix} \mathbf{label}^l_i \\ 1 \end{bmatrix}^T \mathbf{D}_l \begin{bmatrix} \mathbf{label}^r_j \\ 1 \end{bmatrix} \tag{19}$$

where $\mathbf{D}_l$ is a trainable parameter, and $\mathbf{D} \in \mathbb{R}^{\hat{d} \times |\ell| \times \hat{d}}$. $\hat{d}$ is the output dimension of $MLP^{l/r}_{label}$.

2.4 CRF Decoder

In (Zhang et al., 2020), the authors proposed a two-stage framework for constituency parsing, which they proved to have a lower computational cost. Given a sentence x, our goal is to find an optimal tree $\hat{\mathbf{Y}}$. While the previous one-stage method (Stern et al., 2017; Gaddy et al., 2018) tries to parse the optimal tree directly, the two-stage method firstly finds the optimal unlabelled tree.

$$S(x,y) = \sum_{(i,j) \in y} s(i,j) \tag{20}$$

where $S(x, y)$ is the total score of the parsed tree (given x) which is calculated by summing all edge's scores of a legal tree y. y is one of the candidate constituency tree. We put the score under CRF to calculate the conditional probability.

$$p(y|x) = \frac{e^{S(x,y)}}{\sum_{y' \in Tr(x)} e^{S(x,y')}} \tag{21}$$

where $Tr(x)$ is the set of legal trees. Under CRF, the constituency tree is optimized by the CKY algorithm.

$$\hat{\mathbf{Y}} = \arg\max_y p(y|x) \tag{22}$$

The next stage is to identify the label of the optimal unlabelled tree. For each constituent (i, j) of a given tree y and n-length sentence x, the label $\hat{\mathbf{L}}$ of (i, j) is:

$$\hat{\mathbf{L}} = \arg\max_{l \in \ell} s(i,j,l) \tag{23}$$

where ℓ is the set of all possible constituency labels.

According to (Zhang et al., 2020), the complexity of the decoding phase is $\mathcal{O}(n^3 + n|\ell|)$ as the CKY algorithm takes $\mathcal{O}(n^3)$ time complexity and the second stage takes $\mathcal{O}(|\ell|)$ for each edge in y (a constituency tree has $2n - 1$ constituents).

2.5 Loss Function

The loss function of our model consists of 2 parts: the span loss and the label loss. The span loss is calculated by the CRF loss as we try to maximize the conditional probability in equation 21. In other words, given a sentence x and its target tree y, we minimize the $-\log p(y|x)$:

$$L_{span}(x,y) = -S(x,y) + \log \sum_{y' \in Tr(x)} e^{S(x,y')} \tag{24}$$

The second term of equation 24 has an efficiency batchified calculation which is detailed in (Zhang et al., 2020). The label loss L_{label} is computed using the cross entropy function, and the total loss of the model is the sum of two parts:

$$L_{total}(x,y,l) = L_{span}(x,y) + L_{label}(x,y,l) \tag{25}$$

where l is the set of possible labels.

3 Experiment

3.1 Experiment Setup

Data We use the Vietnamese treebank (Nguyen et al., 2009) to train and test our models. The dataset is divided into 3 sets: train, dev, test with 8321, 692, 1388 sentences respectively. The data are pre-processed using the code[4] provided by (Kitaev and Klein, 2018). The root constituents are included in the data, and the function tags are removed due to the purpose of the task. The data are biased as the numbers of 'NP' and 'VP' labels dominate others, which results from the high amount of Noun and Verb words in the treebank. This is heavily affected by Vietnamese grammar, where a simple clause is usually formed from a noun phrase and a verb phrase. Figure 3 and 4 visualize the statistic of constituency labels and POS tags in the dataset.

Parameter choice For the self-attention and label attention layers, we adopt directly the setting of (Mrini et al., 2020) without any tuning except for the number of label attention heads. Following (Mrini et al., 2020), we choose the output dimensions for the MLPs in 2.3 to be 1024 for MLP_{span}'s and 250 MLP_{label}'s. For other parameter settings, we directly follow the choice of (Dozat and Manning, 2017). The token-batch size is 1000, and the training process runs for 100 epochs. The model is evaluated on the performance of the dev set.

Measurement We follow the standard measurement precision (P), recall (R), F-score (F) for evaluation with the help of the `EVALB` tool[2].

Models define Our baseline model is the original CRF model using CharLSTM and BiLSTM-based encoder of (Zhang et al., 2020)[3] with the same settings of the authors. Other models to be compared are listed below:

- *CRF model using pre-trained PhoBERT and BiLSTM encoder*[3]: we examine the impact of our encoder with the LSTM-based one.
- *Berkeley Neural Parser (Benepar) using pre-trained PhoBERT*[4]: we compare our model with the previous method used in (Tran et al., 2020).
- *Our model without using Label Attention Layers*[5]: we consider the contribution of label attention to the parser.

For models using PhoBERT, we re-train them on both base and large versions of PhoBERT to evaluate their influences. Furthermore, we try different setups and use two types of token representations for our LAL to find the optimal choice.

3.2 Results

Table 1 compares our model's scores with other models' on the Vietnam treebank's dev and test set. Overall, our parser (using the large version of PhoBERT) obtains the highest score in all scores (precision, recall, F-score) on both dev and test set. With PhoBERT$_{base}$, the model slightly drops by 0.45 in precision, 0.21 in recall, 0.33 in F-score on the dev set while the numbers are 0.78, 0.48, 0.64 respectively on the test set.

3.3 Evaluation

Evaluation on dev set We conduct the experiments on both versions of PhoBERT (Nguyen and Nguyen, 2020), and we observe that the contribution of Phobert$_{large}$ is greater than PhoBERT$_{base}$. The difference in F-score between the 2 versions ranges from 0.3 to 0.48. When using Character-level LSTM instead of PhoBERT, the scores fall sharply, which proves the essence of the Vietnamese pre-trained model.

With the rich information provided by the encoder and the effectiveness of the Biaffine scorer, our model (w PhoBERT$_{large}$) gives better results in all precision, recall, and F-score on the Vietnamese treebank (Nguyen et al., 2009) as it increases by 1.70% in F-score compared to the Benepar model (w PhoBERT$_{large}$) and 1.78% higher compared to the CRF model (w PhoBERT$_{large}$).

Evaluation on test set On the test set, PhoBERT$_{large}$ continues to outperform

[2] https://nlp.cs.nyu.edu/evalb/

[3] https://github.com/yzhangcs/parser

[4] https://github.com/nikitakit/self-attentive-parser

[5] Our model without LAL refers to removing the *self-attention → LAL* block in our encoder.

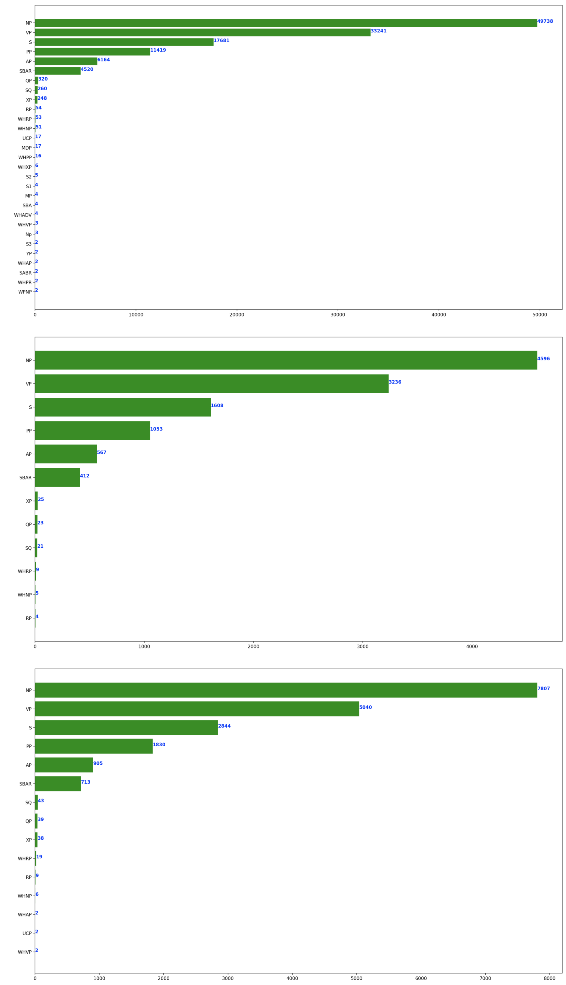


Figure 3: Constituency labels statistic (removed labels whose frequency is 1). Up down: train set, dev set, test set

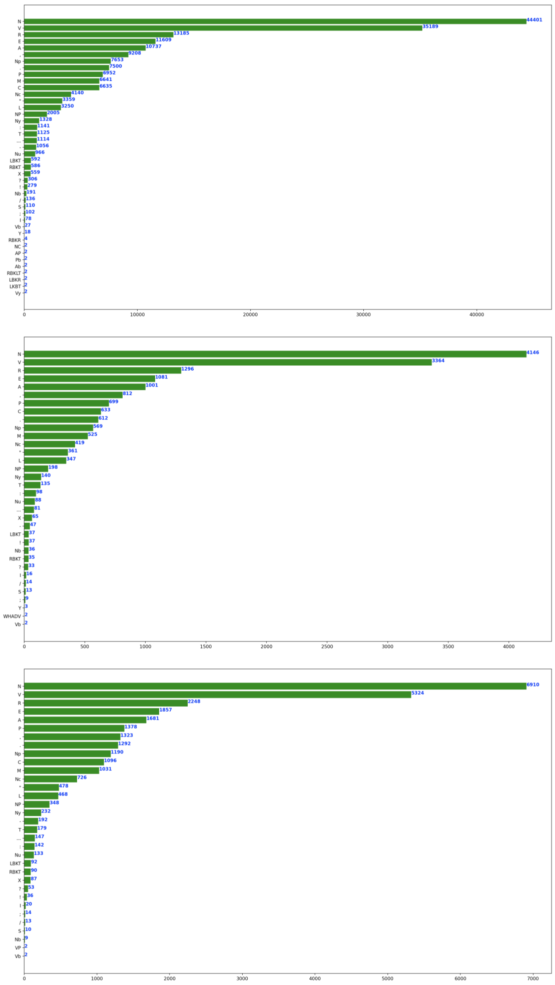


Figure 4: POS tags statistic(removed tags whose frequency is 1). Up down: train set, dev set, test set

PhoBERT$_{base}$ in the Benepar and our models, with 0.51 and 0.64 higher in F-score respectively. However, CRF models give a contrasting result, one of whose explanations can be that the LSTM encoder of the CRF models cannot handle thoroughly massive features gained by PhoBERT$_{large}$.

Our model receives absolute higher results as it achieves a consistent improvement of more than 1 F-score compared to other models. It is clear that the components of our combination benefit from each other and they reconcile to surpass the current architectures.

Impact of Label Attention Layers We examine the importance of the Label Attention mechanism by removing it from the model, and the result drops by 1.42 (the model w PhoBERT$_{large}$) and 0.78 (the model w PhoBERT$_{base}$). Without the LAL, the result is still higher than other tested models.

We make a comparison between different numbers of attention heads in LAL, and table 2 gives the details. In our model defined in 3.1, we use 64 heads for it gives the best F-score on the test set. We test a lower number of heads (32) and surprisingly, the obtained dev set's F-score is better (0.1 higher compared to the 64-head) while the test set's F-score is not significantly lower (0.04 lower compared to the 64-head). Following (Mrini et al., 2020), we choose the last model with 89 heads as there are 89 label heads in Vietnamese treebank and obtain the peak performance on the dev set. Although the 89-head model gives peak performance on the dev set, it requires much more time to train. On the other hand, with the lowest computational cost, the 32-head model still achieves a relatively high F-score on both the dev and test set.

As mentioned in token representations in section 2.2, we use POS embedding instead of word embedding as the content for the *self-attention* → *LAL* block. To make the decision, we trained our model with either the POS or the word embedding whose performances are shown in table 3. Using the word embedding leads to a steep decrease, with 1.65 and 1.68 F-score drops on the dev and test set respectively, which might result from a large amount of the words in the dictionary (the size of the POS's dictionary is only 63).

4 Conclusion

While English and Chinese constituency parsing has achieved significant improvement with different advanced techniques, the task of the Vietnamese language still lacks approaches. We propose an extension of CRF-based model (Zhang et al., 2020) with label attention mechanism (Mrini et al., 2020) to enhance the performance of constituency parsing. This paper takes the advantages of each mechanism to gain a greater impact: pretrained PhoBERT provides knowledge of Vietnamese; self-attention and label attention retrieve the view of words and labels respectively; biaffine attention enhances the scoring framework, and two-stage decoding lowers computational cost. The outcomes of the model are promising with a high F-score, and the idea of combination can be generalized for other languages.

Acknowledgments

This research is supported by research funding from Advanced Program in Computer Science, University of Science, Vietnam National University - Ho Chi Minh City.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing.

David Gaddy, Mitchell Stern, and Dan Klein. 2018. What's going on in neural constituency parsers? an analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 999–1010, New Orleans, Louisiana. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese.

Linh The Nguyen and Dat Quoc Nguyen. 2021. Phonlp: A joint multi-task learning model for vietnamese part-of-speech tagging, named entity recognition and dependency parsing.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 182–185, Suntec, Singapore. Association for Computational Linguistics.

Thi-Phuong-Uyen PHAN, Ngoc-Thanh-Tung HUYNH, Hung-Thinh TRUONG, Tuan-An DAO, and Dien DINH. 2019. Vietnamese span-based constituency parsing with bert embedding. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–7.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Tuan-Vi Tran, Xuan-Thien Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. An empirical study for vietnamese constituency parsing with pre-training.

		Dev set			**Test set**		
		P	R	F	P	R	F
CRF	w CharLSTM	72.41	73.70	73.05	75.78	78.01	76.88
	w pre-trained PhoBERT$_{large}$	80.50	82.10	81.29	82.90	85.19	84.03
	w pre-trained PhoBERT$_{base}$	80.14	81.50	80.81	83.38	85.47	84.41
Benepar	w pre-trained PhoBERT$_{large}$	81.68	81.07	81.37	84.03	85.02	84.52
	w pre-trained PhoBERT$_{base}$	81.04	81.10	81.07	83.65	84.36	84.01
Ours	w PhoBERT$_{large}$	**82.32**	**83.84**	**83.07**	**84.80**	**87.16**	**85.97**
	w PhoBERT$_{base}$	81.87	83.63	82.74	84.02	86.68	85.33
	w PhoBERT$_{large}$ w/o LAL	80.90	82.49	81.69	83.26	85.88	84.55

Table 1: Results table

Number of attention heads	**Dev set**			**Test set**		
	P	R	F	P	R	F
32	82.67	83.70	83.18	84.79	87.09	85.93
64	82.32	**83.84**	83.08	84.80	**87.16**	**85.97**
89	**82.85**	83.65	**83.25**	**85.09**	86.60	85.84

Table 2: Comparison of different number of attention heads of Label Attention

Token representations	**Dev set**			**Test set**		
	P	R	F	P	R	F
word embedding	80.46	82.44	81.43	83.07	85.55	84.29
POS embedding	**82.32**	**83.84**	**83.08**	**84.80**	**87.16**	**85.97**

Table 3: Comparison of different used content embedding

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyu Wang and Kewei Tu. 2020. Second-order neural dependency parsing with message passing and end-to-end training. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 93–99, Suzhou, China. Association for Computational Linguistics.

Kaiyu Yang and Jia Deng. 2020. Strongly incremental constituency parsing with graph neural networks.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural crf constituency parsing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4046–4053. International Joint Conferences on Artificial Intelligence Organization. Main track.

Junru Zhou and Hai Zhao. 2019. Head-Driven Phrase Structure Grammar parsing on Penn Treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2396–2408, Florence, Italy. Association for Computational Linguistics.

Question-Answer Pairing from IM Conversations via Message Merging and Reply-to Prediction

Thamolwan Poopradubsil
Department of Computer Science
Kasetsart University, Thailand
tmw.poopradubsil@gmail.com

Chia-Hui Chang
Department of Computer Science
and Information Engineering
National Central University, Taiwan
chia@csie.ncu.edu.tw

Abstract

Preparing question-answer pairs from conversation logs (chat logs) is often considered a prerequisite for downstream dialogue tasks such as response generation and response selection tasks. In this paper, we study a task called reply-to prediction, which can be used to prepare question-answer pairs. Unlike other works, our data comes from the instant messaging (IM) platform where participants could split long sentences into short utterances and send them in multiple messages. We consider a task called message merging task which aims to determine whether those messages need to be merged or not before generating message pairs for reply-to prediction task. The theory behind this task is similar to, yet different from reply-to prediction task in which this task uses the messages from the same speaker to predict whether these two messages are related or not. We propose a CONTEXT-AOA model to include the context (previous dialogue) as additional input apart from pairwise messages. Our experiments show that our proposed model outperforms both single-turn (pairwise) conversation models and multi-turn (context-aware) conversation models on message merging task and achieves a close performance compares to other multi-turn models on reply-to prediction for manually labeled data and outperforms other models when using heuristic labeled data.

1 Introduction

Preparation of questions-answers from conversation logs (also known as chat logs) between customers and enterprise personnel is important for the development of chatbots. For example, Figure 1 shows a conversation between a customer service staff and a client. The goal here is to find the question-answer pairs such as (d, e), (p, q), from the chat logs such that the question-answer pairs can be used as the chatbot's responses to the corresponding question.

Finding the question-answer pairs is similar to the link prediction task where the goal is to decide whether an incoming message replies to the existing question based on the similarity between messages. Link prediction can be defined either as a reply-to prediction or as a response selection. Both of the problems require message pair similarities to determine which previous or following message is the best-matched candidate question or response. Similar to conversation disentanglement, earlier works on response selection only use the last message in the context for matching with the candidate response message from different utterances (Wang et al., 2013). However, human does not give the response entirely based on a link prediction task. Real-world chat logs are multi-turn conversations, thus including the context as another input is necessary in order to allow a model to remember what has been discussed so far. Recent works show that including the multi-turn conversation improves the response selection's performance, e.g. (Zhou et al., 2016; Zhou et al., 2018; Wu et al., 2019).

In this work, we focus on online instant messages between customers and service personnel. The major problem is that a complete message (either the question or the answer) may be split into several ut-

Data Set			Manual		Heuristic	
ID	Author	Messages	ID_M	RT	ID_H	RT
a	Client	need support with an open ticket	0		0	
b	QNAP	paoyang	1		1	0
c	QNAP	Hi, how are you doing				
d	QNAP	what is the ticket #?	2	0		
e	Client	#FNA-202-51465	3	2	2	1
f	Client	it is for 3M				
g	Client	they are very unhappy with the ...				
h	Client	ive been on hold on the phone ...				
i	QNAP	I am working on that ticket.	4	3	3	2
j	QNAP	They are using an embedded ...				
k	QNAP	According to the tech who worked on it ...				
l	Client	have you let them know ...	5	4	4	3
n	Client	so will it not work at all with their server?				
m	QNAP	so this seems like an issue with ...	6		5	4
o	QNAP	It seems to be closely related to ...				
p	Client	So when should I tell ...	7	6	6	5
q	QNAP	I will respond back to them but ...	8	7	7	6
r	Client	Okay, I will let them know of our conversation.	9	8	8	7
s	Client	Thank you.				

Figure 1: A real-world conversation between QNAP customer service and client.

terances. For example, e, f, g and h in Figure 1 together make a complete answer to question d, therefore these four messages should be merged together into one utterance. However, we cannot directly merge all adjacent utterances from the same speaker since each message might be either an incomplete message related to its previous utterance or a complete message on its own. For example, c and d should not be merged since message d alone makes a complete response to message a while message c asks about something unrelated to the question in message a.

To deal with this kind of dataset, we propose a two-stage pipeline for the question-answer pair preparation. We distinguish the task of extracting question-answer pairs into two subtasks: message merging task and reply-to prediction task in order to solve the problem we mentioned previously. We first regard both subtasks as a sentence pair classification task within a single-turn conversation setting. Three neural networks models based on GloVe word embedding (including CNN+LSTM, LSTM with dual attention, and attention over attention (AOA) (Huang et al., 2018)) as well as two BERT (Devlin et al., 2019) based models (BERT sentence pair classification and the combination of BERT-SPC and AOA) are proposed. However, the best performance only achieves 0.763 and 0.794 accuracy (0.443 and 0.498 F1) on the message merging task and reply-to prediction task respectively.

To further improve the performance, we consider including context (previous dialogue) as additional input apart from only pairwise messages. With the application of AOA over any pairs of the context and two messages, we show a significant improvement over the previous models for message merging (0.964 accuracy and 0.887 F1) and even outperform existing models on reply-to prediction task when using heuristic labeled data (0.956 accuracy and 0.823 F1).

The rest of the paper is organized as follows: First, we give a definition of the two tasks (message merging task and reply-to prediction task)r. Then, we introduce the dataset used in this paper and describe the preparation process. Next, we describe our proposed model (CONTEXT-AOA). Finally, we show the question-answer pair's preparation model for both tasks and compare them with the results from existing models.

2 Related Work

2.1 Conversation Disentanglement

The idea of treating the disentanglement task as a two-stage problem which has later been the major idea for most existing studies on this task has been proposed by Elsner and Charniak (2008). The first stage is a binary classification task where the main goal is to mark each message pair as alike or different while the second stage utilizes the results from the first stage to cluster the utterances into each conversation thread (same thread prediction).

Earlier approaches to the first stage or link (reply-to) prediction are based on a statistical classifier with the use of linguistic features in order to calculate the similarity of each message pair, e.g. (Elsner and Charniak, 2008; Elsner and Charniak, 2010). Recent approaches use neural models to learn abstract linguistic features. Mehri and Carenini (2017) use handcraft features along with the probability of being the next utterance which is predicted from a pre-trained LSTM model to train the reply classifier. Jiang et al. (2018) introduced Siamese Hierarchical CNN (SHCNN) to investigate how message similarity could be estimated. SHCNN uses hierarchical CNN to capture both low-level and high-level message meanings of each message. The interaction between two message representations which is computed using element-wise absolute difference methods is then used along with the handcraft features to estimate the similarity between two messages.

2.2 Single-turn vs. Multi-turn Response Selection

Earlier works on response selection tasks are only based on pairwise message comparison. Lu and Li (2013) proposed a DNN-based matching model for short text response selection by extracting the local pairwise relations on a low level with patches and sending them to the DNN layers. Hu et al. (2014) utilize deep CNN in order to capture the interaction between message and response.

However, using only pairwise messages might not be enough to solve this task, Lowe et al. (2015) introduced the task called multi-turn conversation response selection pair requires matching between a response and a conversation context (previous utterances) instead of a single previous message. They applied LSTM on the concatenated utterances (context) and a response message to perform context-response matching on a word-level context vector. Later, the work by Zhou et al. (2016) extended this idea by performing context-response matching on noa t only the general word level context vector but also the utterance level context vector. However, ignoring the relationships between the utterances (context) when concatenating them together (Lowe et al., 2015) or converting the whole context to a vector without enough supervision from responses (Zhou et al., 2016) might be the cause of some information loss which affects the model performance. To address this problem, the work by Wu et al. (2017) proposed a sequential convolutional network (SCN) that matches a response with each utterance in the context with a CNN and then accumulated the matching vectors in the utterances' temporal order to model their relationships. Another work from Wu et al. (2019) replaced the convolution neural network (CNN) with an attention layer to extract the interaction between each utterance and response.

2.3 Message Pair Classification

To build a better model for reply prediction tasks, we also refer to other tasks that accept two messages as input such as aspect-based sentiment analysis (ABSA) and natural language inference (NLI).

Aspect-based sentiment analysis aims to determine the sentiment polarity of a given sentence and aspect. Many models and methods have been proposed from traditional machine learning methods (Kiritchenko et al., 2014) to deep learning models (Liu et al., 2020). For example, Wang et al. (2016) proposed an attention-based LSTM network for aspect-based sentiment classification. Huang et al. (2018) introduced an attention-over-attention (AOA) neural network to capture the interaction between aspects and context sentences, which outperform LSTM-based architectures. However, one drawback of LSTM-based architectures is that their training process is time-consuming. To address this problem, Xue and Li (2018) introduced GCAE model where its main components are CNN and gating mechanisms to reduce the number of training parameters and speed up the procedure.

On the other hand, the task of natural language inference is to determine if one given statement (a premise) semantically entails another given statement (a hypothesis). For example, Parikh et al. (2016) proposed "Decomposable Attention Model" which uses a shared sentence representation with fewer parameters and mutual attention mechanism to build a model with high performance.

3 Problem Definition and Dataset

Instant messaging (IM) is a type of online chat which offers real-time text-based communication in which two persons participate in a single conversation over their computers or mobile devices within an Internet-based chatroom. This type of online chat is commonly used in many business companies in order to support their clients. Companies often store their conversation logs between customer service staff and clients in order to be used in the future to improve their service and customer experiences.

The main goal of this work is to predict whether one message is a reply-to message to a previous message. However, unlike messages from other sources such as email or Reddit forums, messages from instant messaging are generally short to enable quick response. Moreover, participants could split long sentences into short utterances and send them in multiple messages. Therefore, some messages need to be merged before generating reply-to message pairs.

The overview of the training phase is outlined as follows, where we build a merging model to decide whether two messages from the same speaker need to be merged and a reply-to model to decide whether two messages from two speakers are a reply-to message pair.

- **Message merging task** aims to determine whether a given message pair from the same speaker should be merged or not. Most continuous messages from the same speaker could be merged, however, some messages should not be merged when they deliver different intentions. As shown in Figure 1, message e to h can to be merged in order to make a complete response to message d, while message d should not be merged with message b and c because they serve different purposes.

- **Reply-to prediction task** is to determine whether a given message pair from different speakers is a question-answer pair or not. For example, the message $ID_M = 2$ is a response to message $ID_M = 0$ (positive) while the merged message $ID_M = 1$ is not a response to message message $ID_M = 0$ (negative).

3.1 Training Data Preparation

An in-house QNAP customer service dialogue dataset contains conversation logs between two participants: a QNAP client and a customer service staff. QNAP customer service staff may reply to previous messages for troubleshooting or ask questions to clarify the problem while assisting the clients. The chatlog disentanglement task here is to find direct reply messages for question-answer pairing. A total of 1,860 conversations with an average of 31.7 messages per conversation are collected as our dataset.

Manual Annotation

We randomly select 60 conversations and ask three annotators to label these conversations. Two of them are asked to label the 60 conversations while the third annotator is asked to make the final decision on the inconsistent labels. For the message merging task, we only focus on the message pairs where both two messages are from the same speaker. The merged message is renumbered with message ID ID_M. For the reply-to prediction task, the annotators will focus on the merged messages and mark the current message as a response to which previous ID_M from the other speaker.

Message Pairs Preparation Given the annotated messages, we can pair messages that need to be merged as a positive example and independent messages as a negative example. Since each message (2,136 messages for the message merging task and 1,366 messages for the reply-to prediction task) can be paired with all of its previous messages in the same conversation, the number of negative examples is much larger than that of positive examples. Thus, the kappa value from the two annotators is quite high

(0.876 and 0.990) for the merging and reply-to tasks, respectively.

Dataset	Chat Log		
Conversations	60		1,800
Labeling	Manual	Heuristic	Heuristic
# Messages	2,136	2,136	56,792
Pos. Pairs	770	1,082	29,437
Neg. Pairs	3,679	-	-
# Pairs	4,449	-	-
# Merged Msg	1,366	1,054	27,355
Reply-to	743	753	19,188
Non-Reply	4,032	3,407	86,290
# Pairs	4,775	4,160	105,478

Table 1: Training and Testing data in chat log

With the third annotators, we get a total of 770 positive pairs and 743 reply-to pairs. To deal with imbalanced data, we restrict the maximum number of negative message pairs for each message. That is to say, we conduct negative message pairs down sampling. For the message merging task, we set the maximum number of negative message pairs equal to 2 which means we randomly choose at most 2 negative pairs from all possible negative pairs for each message, and for the reply-to prediction task, we set the maximum number of negative message pairs equal to 4. Overall, we have 3,679 negative pairs for the message merging task and 4,032 non-reply pairs for the reply-to prediction task from the 60 conversations via manual labeling.

Heuristic Labeling

The heuristic labeling data is prepared by merging all consecutive utterances from the same speaker into one message under the assumption that a client service staff's message following the customer's question in the previous turn is a reply to the question and using this assumption to annotate all conversations automatically. As shown in Figure 1, adjacent messages from the same speaker are merged into one message and are renumbered with a new ID, ID_H. Since the heuristic labeling always merges the adjacent messages from the same speaker, there are no negative examples. We then generate message pairs $x = (i, j)$ with $i < j$, where x is a positive reply-to example if j equals $i + 1$, otherwise a negative (Non-Reply) example. Similar to the manual labeling process, we conduct negative example down sampling to prepare pairs. Finally, we generate 19,188 reply-to pairs and 86,290 non-reply pairs from 1,800 conversations.

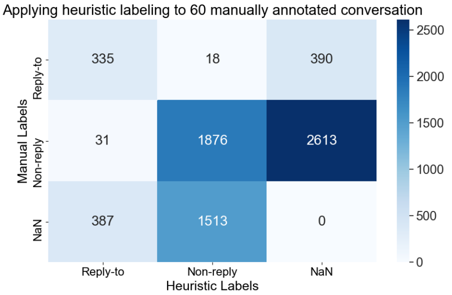


Figure 2: Applying heuristic labeling to 60 manually annotated conversations.

To see how effective the heuristic labeling is, we take the 60 manually annotated conversations as golden answers and compare them with the heuristic labeling result for performance evaluation. Since heuristic labeling merges all consecutive messages by the same speaker, the merge ratio (0.51=1082/2136) is higher than that of manual labeling (0.36=770/2136) (see Table 1). Due to the difference in the message merging step, some messages do not have corresponding matches in the other labeling method. Therefore, a third class "NaN" is used to denote message pairs that do not have corresponding matches as shown in Figure 2. Excluding unmatched pairs, the heuristic labeling has a 0.92 (=335/(335+31)) precision and 0.95 (=335/(335+18)) recall.

4 Context-Aware Message Pair Classification Models

Both the message merging and reply-to prediction models can be regarded as sentence pair classification models. However, the existing models which use only question-answer pair as input might not be enough for a task such as chat log (conversation) disentanglement. Inspired by the works on response selection task and AOA (Huang et al., 2018), we propose a model which includes the previous dialogue (context) as an input in addition to the question-answer pair. The overall model structure is as shown in Figure 3.

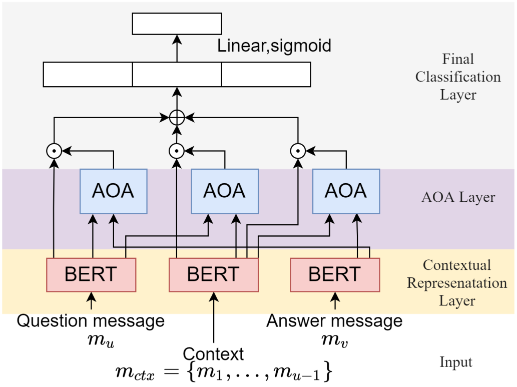


Figure 3: Context Attention-over-Attention BERT model structure ($u < v$).

Contextual Representation Layer

For each message pair (m_u,m_v) from a chat log C, we also include its context $m_{ctx} = \{m_1, ..., m_{u-1}\}$ as part of the input. That is to say, each training example is a triplet tuple $\mathbf{x} = (m_{ctx}, m_u, m_v)$. If m_v is a reply to m_u, $\mathbf{x}$ is considered a positive example, otherwise it is a negative example. We then apply shared BERT embedding to get the representation $h \in R^{l \times d}$ of each message m respectively, where l is the number of tokens after BERT word piece subword segmentation and d (=768) is the dimension size of BERT embedding.

Attention-over-Attention (AOA) Layer

Given two message representation $M_1 \in R^{n \times d}$ and $M_2 \in R^{m \times d}$, AOA first calculates a pair-wise interaction matrix $I = M_1 \cdot M_2^T$, where the value of each entry I_{ij} represents the correlation of a word pair among the two input messages. Next, two matrix column-wise softmax, $\alpha \in R^{n \times m}$ and row-wise softmax, $\beta \in R^{n \times m}$ are computed as follows.

$$\alpha_{ij} = \frac{exp(I_{ij})}{\sum_{k=1}^{n} exp(I_{kj})}, \beta_{ij} = \frac{exp(I_{ij})}{\sum_{k=1}^{m} exp(I_{ik})}, \quad (1)$$

The idea of AOA is to use the averaged attention weight $\overline{\beta} \in \mathbb{R}^m$ for the computation of output feature vector $\gamma \in \mathbb{R}^n$, where

$$\overline{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \beta_{ij}, \quad (2)$$

and the output of the attention-over-attention layer structure is computed by using $\overline{\beta}_j$ as a weight for each α_j:

$$AOA(M_1, M_2) = \alpha \cdot \overline{\beta}^{\mathsf{T}}. \quad (3)$$

Suppose the output of BERT embedding for the training example is denoted as M_u, M_v, and M_{ctx}. We then pair these embedding and apply attention-over-attention over three pairs to obtain $AOA(M_{ctx}, M_u)$, $AOA(M_{ctx}, M_v)$ and $AOA(M_u, M_v)$.

Final Classification Layer

Next, we use $AOA(M_1, M_2)$ for calculating the attention-weighted representations of each input pair.

$$r(M_u, M_v) = M_u^{\mathsf{T}} \cdot AOA(M_u, M_v)$$
$$r(M_{ctx}, M_u) = M_{ctx}^{\mathsf{T}} \cdot AOA(M_{ctx}, M_u)$$
$$r(M_{ctx}, M_v) = M_{ctx}^{\mathsf{T}} \cdot AOA(M_{ctx}, M_v)$$

Finally, we concatenate all the attention-weighted representations to the prediction layer, i.e. $\mathbf{p_o} = r(M_u, M_v) \oplus r(M_{ctx}, M_v) \oplus r(M_{ctx}, M_u)$ by Eq. 4.

$$P(y|x) = \sigma(\mathbf{w} \cdot \mathbf{p_o} + b_o) \quad (4)$$

5 Experiments

During testing time, we are given a chat log that is not labeled. We simply apply the merging model and reply-to prediction model in order as shown in Figure 4:

1. **Message pairing (Same speaker)**: We first pair the messages from the same speaker based on the trained merging models to determine whether these message pairs should be merged or not. The messages will be merged according to the output from the merging model. We then update the chat log file by replacing the message pairs that need to be merged with the merged messages.

2. **Message pairing (Different speaker)**: We then pair reply-to message pairs using the chat log we obtained in the previous step. Unlike the

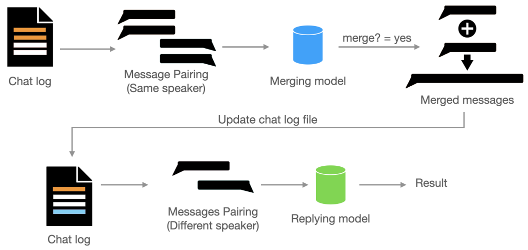


Figure 4: The testing phase of reply message prediction model.

first message pairing step, in this step, we focus on the message pairs from a different speaker. These pairs are then given to one of the reply-to prediction models to decide whether they are correct reply-to message pairs or not.

5.1 Experimental Setup

We divide the manually labeled data into 5-fold and use either 4-fold (48 conversations) out of the 60 conversations as training data to build the prediction models (both merging model and reply-to model for manually labeled examples). The models are tested on the remaining 12 conversations. The process is repeated five times and the result is averaged to obtain the final result. For heuristically merged and labeled examples from 1,800 conversations, we train the reply-to prediction models and test on all 60 conversations to compare the performance.

We implement three GloVe-based neural network models and two BERT models for performance comparison.

GloVe-based Models

A typical neural network model consists of an embedding layer for word representation, a hidden layer such as mutual attention for message representation, and an output layer for prediction. For the embedding layer, we adopt a pre-trained GloVe (Pennington et al., 2014) word embedding matrix from the Common Crawl dataset (42B tokens), which contains a case-sensitive vocabulary of size 1.9 million. We consider three models for message representation. The first one is GCNN-LSTM, the second is LSTM with dual attention, and the third is Attention-over-Attention (AOA) model.

- **GCNN-LSTM Representation** We use Convolutional Neural Networks (CNN) for feature extraction with Gated Linear Unit (GLU) proposed in (Dauphin et al., 2017) to control which information flows in the network. To deal with word sequence, we adopt a BiLSTM layer to capture the message information. The outputs from the BiLSTM layer are passed through two fully connected layers to make the prediction.

- **LSTM Dual Attention Model** Inspired by the power of the attention mechanism, the second model we proposed is BiLSTM with dual attention where we can exploit the attention mechanism to generate a representation for m_1 based on the content of m_2.

- **Attention-over-Attention (AOA) model** The above two models only focus on message representation. Therefore, we exploit the idea of capturing the interaction between one message to another message given the hidden semantic representations of the two messages generated by BiLSTMs with AOA. (Huang et al., 2018).

The pre-trained GloVe word embedding has a dimension size of 300. The hidden layers in BiLSTMs are 128, 128, and 300 for GCNN+LSTM model, LSTM+DualAtt model, and AOA model respectively, the number of kernels used in CNN is 128 with the kernel size equal to 5. The batch size used in the traditional deep learning model is 128 and the maximum epoch and initial learning rate are set to 40 and $1 * 10^{-3}$.

BERT-based Models

Different from context-free models, which generate a fixed word embedding representation for each word in the vocabulary, BERT is able to give a context-dependent representation of the words. Consequently, we use the BERT model released by Google and fine-tune it for the message merging/reply-to prediction task.

Given two input messages m_1 (with length n) and m_2 (with length m), we employ BERT component with L transformer layers to calculate the corresponding contextualized representations with input of the form $([CLS], m_1, [SEP], m_2)$.

- **BERT-SPC** The basic BERT sentence pair classification (BERT-SPC) model takes the output of [CLS] token as the prediction layer input.
- **BERT-SPC-AOA** We exploit the idea of Attention-Over-Attention model to further improve the BERT-SPC model by concatenating the output from AOA with [CLS] output as the input to the prediction layer.

For BERT based model, the batch size is 16 and 8 for CONTEXT-AOA model. The maximum epoch and initial learning rate are 6 and $2 * 10^{-5}$, respectively. The optimizer used in all models is Adam with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. All models are trained on GeForce GTX1080Ti 10GB GPU.

5.2 Performance Comparison

Table 2 and 3 show the performance comparison of the proposed CONTEXT-AOA model with both single-turn models based on message pair similarity and multi-turn chatlog disentanglement models with additional context.

Message Merging Task

For message merging task, single-turn approaches with only two message input exhibit limited performance. The highest F1 score of these models is only 0.443 F1, which is achieved by BERT-SPC-AOA model as shown in Table 2, While all multi-turn approaches including (Lowe et al., 2015), (Zhou et al., 2016), and (Wu et al., 2019) have significant improvement over single turn approaches. The proposed CONTEXT-AOA model achieves the best 0.887 F1 and 0.964 accuracy.

	QNAP: Message Merging Task		
	Model	F1	Acc
Single-turn	SHCNN (Jiang et al., 2018)	0.266	0.763
	GCNN+LSTM	0.271	0.731
	LSTM+DualAtt	0.254	0.680
	AOA (Huang et al., 2018)	0.333	0.516
	BERT-SPC	0.374	0.734
	BERT-SPC-AOA	0.443	0.731
Multi-turn	LSTM (Lowe et al., 2015)	0.859	0.958
	MultiView (Zhou et al., 2016)	0.841	0.944
	SAN (Wu et al., 2019)	0.851	0.948
	CONTEXT-AOA	**0.887**	**0.964**

Table 2: QNAP chat log: Message Merging Task

Reply-to Prediction Task

For reply-to prediction task, we see a similar result. The proposed CONTEXT-AOA model yields 0.800 F1 and 0.944 accuracy, while the highest F1 score of single-turn models is 0.498 (by BERT-SPC-AOA model) as shown in the “Manual” column of Table 3. The experimental results demonstrate that one cannot neglect the relationship between previous messages (context) and the question-answer pair. Including previous messages as additional input significantly improves the performance for both subtasks on manually labeled examples.

For the result shown in the “Heuristic” column of Table 3, we train the models using all of the heuristically labeled data as training data and test on manually labeled examples where we divided the testing data into 5 folds and train the models similar to what we’ve done with the experiments on manually labeled examples.

Interestingly, the heuristically labeled data provides better performance for most of the multi-turn reply-to prediction models. The result may be attributed to a large amount of training data even though the heuristic labeling rule does not always generate the correct labeled data. This might also be the cause of an unstable performance for several models.

Figure 5 shows the performance of multi-turn conversation task models on heuristic labeled data in regard to the size of training data. We find that all models exhibit a steep slope the training data size is lower than 5%. Moreover, using 25%-50% of heuristic labeled data to train the models can significantly outperform the full size of manually labeled data, which is about 5% of the heuristic labeled data.

6 Conclusion

This paper addresses the problem of question-answer pairs preparation from two participants’ online chat logs. The major problem with this kind of data is that a complete message may be split into several utterances, therefore additional task (message merging task) is required to merge some of these utterances together before forming the question-answer pairs. To extract question-answer pairs from chat logs, we perform reply-to prediction task on merged messages in order to identify the

	QNAP: Reply-to prediction				
	Data Set	Manual		Heuristic	
	Model	F1	Acc	F1	Acc
Single-turn	SHCNN (Jiang et al., 2018)	0.429	0.722	0.432	0.702
	GCNN+LSTM	0.390	0.680	0.441	0.694
	LSTM+DualAtt	0.403	0.706	0.461	0.710
	AOA (Huang et al., 2018)	0.449	0.754	0.394	0.709
	BERT-SPC	shown1	0.736	0.413	0.665
	BERT-SPC-AOA	0.498	0.794	0.410	0.717
Multi-turn	LSTM (Lowe et al., 2015)	0.750	0.932	0.796	0.946
	MultiView (Zhou et al., 2016)	0.802	0.943	0.820	0.950
	SAN (Wu et al., 2019)	**0.814**	**0.948**	0.811	0.952
	CONTEXT-AOA	0.800	0.944	**0.823**	**0.956**

Table 3: QNAP chat log: Reply-to prediction

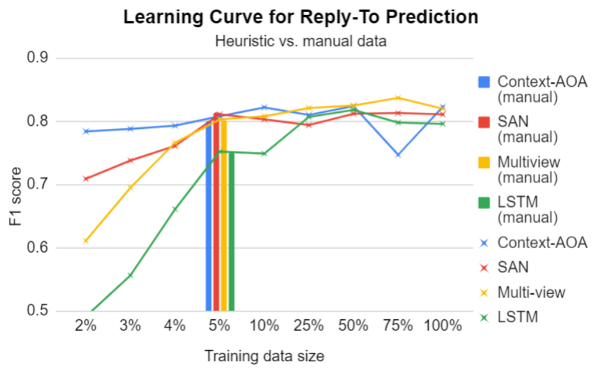


Figure 5: Learning curve for reply-to prediction task.

correct question-answer pairs. In terms of model design, we propose a context-aware AOA model which utilizes the idea of Attention-over-attention models to capture the relationship between context, question, and answer message.

Experimental results on both message merging and reply-to prediction tasks show that allowing the model to gain access to the context significantly improves the performance on both tasks. Our proposed CONTEXT-AOA model outperforms the existing models on message merging task and achieves comparable performance on reply-to prediction task for manually labeled data. In addition to manually labeled data, we also conduct experiments on heuristically labeled data where our proposed model outperforms the existing models and the result demonstrates that more training data may further improve the problem.

Acknowledgments

The research is partially supported by Ministry of Science and Technology, Taiwan under grant MOST109-2221-E-008-060-MY3.

References

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio, June. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2042–2050. Curran Associates, Inc.

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. *CoRR*, abs/1804.06536.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana, June. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August. Association for Computational Linguistics.

H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September. Association for Computational Linguistics.

Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1367–1375. Curran Associates, Inc.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA, October. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada, July. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197, March.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July. Association for Computational Linguistics.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas, November. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia, July. Association for Computational Linguistics.

The Representation of Discontinuity and the Correspondence Principle

Ratna Nirupama
Indian Institute of Technology Hyderabad
Sangareddy
Telangana, India
la19resch11006@iith.ac.in

Prakash Mondal
Indian Institute of Technology Hyderabad
Sangareddy
Telangana, India
prakashmondal@la.iith.ac.in

Abstract

Discontinuity is a nearly universal phenomenon observed in natural languages. Several approaches have been proposed so far by different grammar formalisms but they are widely regarded as distinct approaches owing to their theoretical motivations. This paper proposes the *correspondence principle* which will enable the representation of discontinuity by way of the unification of the representations of linguistic structures in three grammar formalisms: Phrase Structure Grammar (PSG), Dependency Grammar (DG), Categorial Grammar (CG). The goal is not to unify PSG, DG and CG, but rather to sketch out a way of representing discontinuity by uniting constituency relations (as in PSG), head-dependent relations (as in DG) and functor-argument relations (as in CG) for the encoding of discontinuous expressions in natural languages. The implications for natural language syntax and computational linguistics will be offered towards the end of the paper.

1 Introduction

Syntactic discontinuity is a grammatical phenomenon in which a constituent of a sentence is split into two (or more) parts because of the insertion of an element which is not a part of the constituent. The evidence for discontinuity is frequently found in languages with relatively free word-order such as Turkish, Russian, Japanese, Croatian, German, Tamil, Warlpiri etc. In the sentence from Malayalam below, the predicate and its argument are not contiguous as per the linear order because the subject is located between them, indicating a case of discontinuity.

i. *Kaṇṭu kuṭṭi* *aanaye*. (Falk, 2001:19)
saw child.NOM elephant.ACC
'The child saw the elephant.'
[NOM=nominative case marking;
ACC=accusative case marking]

It is also observed in rigid word order languages such as English, but is limited to long-distance dependencies such as topicalisation, long-distance *Wh*-questions etc. Various theories of grammar have accounted for discontinuity in natural languages (both rigid and free word order) in different ways, as per their theoretical motivations. For example, PSG has rules and analyses syntactic structures only in terms of constituents/phrases, making it well-nigh impossible to account for discontinuous constituents. DG and related formalisms accommodate discontinuous constituents by analysing expressions on the basis of head-dependent relations. The paper introduces *the correspondence principle* which will help in the unification of the representations of linguistic structures in these grammar formalisms for some cases of discontinuity. Firstly, a brief introduction to each of the grammar formalisms is given followed by a mention of the previous approaches towards discontinuity. This is followed by an illustration of *the correspondence principle* which will help in the derivations, considering one discontinuous sentence from Croatian language. Finally, a brief conclusion is provided towards the end of the paper.

2 The Three Grammar Formalisms: PSG, DG and CG

2.1 PSG

Analysing sentences as constituents is central to the PSG formalism which was first put forth by Noam Chomsky in his book ‘Syntactic Structures’ (1957) and later developed in the Extended Standard Theory, The Revised Extended Standard Theory, Government/Binding theory and the Minimalist Program. They all exhibit certain common characteristics: all syntactic representations are analysed as phrases and depicted using trees; the grammatical functions are derived from the constituent structures; the configuration of the subject is higher and external; certain operations called transformations (hence transformational grammar) on an existing constituent structure change it into a similar, but not identical, constituent structure called the 'surface structure'. Thus, the traditional PSG greatly emphasizes the interdependence of grammatical relations, thematic roles and constituency. Rewriting rules based on PSG trees specify the manner in which each of the word/phrase is combined to form constituents. For instance, in the discontinuous sentence (i), PSG depicts the predicate *Kaṇṭu* and its argument *aanaye* as V and NP respectively (S → V NP NP). However, *Kaṇṭu* and *aanaye* are considered to form a single VP constituent (VP → V NP) if it were a continuous sentence. Thus, any word order variations are encoded in the rewriting rules based on PSG trees depicting the dominance and linear relations among various constituents (Gazdar, 1983; Chomsky, 1995; Newmeyer, 2001).

2.2 DG

DG is a descriptive tradition in linguistics that can be traced back to Panini and was later developed by the French linguist Lucien Tesnière (1959). It analyses sentences in terms of head and dependent relations, motivated by the grammatical functions. A DG can be specified by a 4-tuple: DG = <V_N, V_T, D, R> where V_N is the set of auxiliary/non-terminal items (syntactic categories), V_T is the set of terminal items (actual words realized from syntactic categories), D is the set of dependency rules and R is the initial symbol at the root of the tree (that is, R∈ V_N). A dependency rule in D is a statement consisting of one auxiliary element functioning as the governing element or head (I) and any finite number of auxiliary elements as the dependents. There are two important rules in D: Rule 1: I(D_1,…,D_m * D_i,…,D_k) (i, m, k ≥ 0 ; not always i=m=k) ; Rule 2: I (*); ‘I’ is the governing element and indicates the presence of only one independent category; D_1,…,D_m represents the dependent categories towards the left of the root word/head; D_i,….,D_k represents the dependent categories towards the right of the head word; *m* and *k* are the number of dependents on the left and right of the head word respectively. The asterisk ‘*’ indicates the location of 'I' in the linear order of words. As per this rule, the valence of ‘I’ will be the total number of dependent elements i.e. ‘(m+k)’. Therefore, the terminal elements, the non-terminal elements, the correspondences (dependency functions) which exist between them, and the rules constitute the core of the dependency theory (Hays, 1964; Gaifman, 1965; de Marneffe and Nivre, 2019). Accordingly, the rule in D for the sentence *i* is V(*NP, NP) which will be realized as *Kaṇṭu* (**kuṭṭi*, *aanaye*). This indicates that V *Kaṇṭu* is the head and NPs *kuṭṭi* and *aanaye* are its dependents. The central idea is that in a sentence, all except one word (‘the root’) depend on another word. A dependent Y depends on a head X when Y is usually optional with respect to X, and/or X selects Y, and/or Y agrees or is governed by X, and/or the linear position of Y is with reference to that of X (de Marneffe and Nivre, 2019: 203).

2.3 CG

Thirdly, CG is a context-free grammar formalism first defined by the logician Kazimierz Ajdukiewicz (1935). The notion of ‘category’ and analysis of sentences in terms of functor-argument relations constitute the core idea of this formalism. Words are assigned a category in terms of N and S, based on their combining properties (Steedman, 1992, 2014). The widely used ‘slash’ notations for directional categories were pioneered by Bar-Hillel (1953) and Lambek (1958). Lambek’s notation uses a forward slash ‘/’ to indicate an argument on the right and a backward slash ‘\’ to indicate an argument on the left. It needs to be emphasized that for the CG analyses in this paper, the standard Lambek notation of functor-argument relations (by using only the categories N and S) has been adopted. Given this basic understanding, the next section makes a brief note on the solutions proposed so far to solve the problem of discontinuity. A CG can be defined by a 4-tuple: CG = <V, C, R, F>. Here, V is the set of all lexical items in a language; C is the set of primitive categories ({N, S}); R is the set of functional composition rules for the generation of categories for lexical items. It specifies the process of generation of category of any given lexical item. F is a function that maps each lexical item (LI) to its set of categories (each element of V is mapped to its corresponding element(s) which can be a set of primitive/atomic categories from the set C and/or categories derived by means of R), whose form is: F(LI) = {$C_1,\ldots,C_n$}. For instance, for the Malayalam sentence (i), V = {*Kaṇṭu*, *aanaye*, *kuṭṭi*}; C = {N,S}; as per the definition of R and F, the category of *Kaṇṭu* is '(S/N)/N' (with '/' indicating the argument is to the right) and 'N' is for *aanaye* and *kuṭṭi*. In Step 1 of the CG derivation, *Kaṇṭu* is the functor and its argument *aanaye* is to its right. Here, the cancellation of 'N' results in the output '(S/N)' (*Kaṇṭu aanaye*). In Step 2, (S/N) is the functor and N (*kuṭṭi*) is the argument. Here, 'N' is cancelled out, which results in the final output S. We shall now look into the previous approaches to discontinuity proposed so far in the literature.

3 Previous Approaches to Discontinuity

Questioning the validity of a universal constituent, several linguists have proposed alternative approaches towards discontinuity.

3.1 Characterisation of Non-Configurational Languages

Hale's (1982, 1983) work on Australian languages such as Warlpiri, Navajo and Dixon's (1972, 1977) work on Dyirbal and Yidiny provided rich evidence for the existence of discontinuity in natural languages. Hale associated three key properties with 'non-configurational languages':

(i) free word order
(ii) the use of syntactically discontinuous expressions, and
(iii) the extensive use of the null anaphora (an argument such as subject and object that is not represented by an overt nominal expression in the phrase structure).

This is because the syntactic nature of these languages is not the same as that of more familiar languages which admit of analyses in terms of phrase structure constituency (the structure of a clause, configurations of NPs and VPs), subordination, *wh*-movement and extraposition (Nordlinger, 2014). Austin and Bresnan's (1996) claim of Warlpiri phrase structure as flat and characterised by free base-generation of elements is another approach towards discontinuity.

3.2 Phenogrammatical Structure

Dowty (1996) makes two important assumptions. First, he proposes a 'minimalist theory of syntax' to describe various discontinuous syntactic phenomena by taking linear structure as the norm rather than hierarchical structure, that is, 'a clause or a group of words is *only* a string'. Second, some words and constituents are more tightly bound (attached) to adjacent words than others. The linear structures/representations of expressions are treated as unordered lists.

3.3 Sequence Union Operation/Shuffle

Donohue and Sag's (1999) adopted Reape's (1996) 'sequence union operation' or 'shuffle'. The sequence union of two lists l_1 = <a,b> and l_2 = <c,d> is the list l_3 iff each of the elements in l_1 and l_2 is present in l_3 and the original order of the elements in l_1 and l_2 is preserved. For example, the sequence union of l_1 and l_2 is any of the following lists/sequences: <a,b,c,d>, <a,c,d,b>, <a,c,b,d>, <c,d,a,b>, <c,a,d,b>, <c,a,b,d> but not <b,a,c,d>, <a,b,d,c> etc. This allows discontinuous elements to intervene in the linear order of a constituent, thus accounting for discontinuity.

3.4 Tangled Trees

McCawley's tangled trees (McCawley, 1982; Iwakura, 1988; Blevins, 1990) relax the *no-crossing constraint* and the *single mother condition* of the standard PSG trees to account for discontinuity.

3.5 Parallel Merge

Citko's (2011) 'parallel merge' relaxes the *single root/mother condition* to linearize multidominant structures, thus accounting for discontinuous structures.

3.6 Encoding Constituents in terms of Dependency Relations

These include Barry and Pickering's (1990, 1993) 'dependency constituent' linking dependency relations with constituent relations, the formulation of 'subtrees' from DG trees by Hays (1964), Gaifman's (1965) formulation of weak equivalence between 'parenthetical expressions' of PSG trees and dependency graphs.

However, these are limited to showing the correspondences between PSG and DG. This paper goes beyond this and attempts to show a way of unifying CG functor-argument relations with DG head-dependent relations and constituency rules in PSG. Now we shall introduce *the correspondence principle* and the motivations behind this proposal.

4 The Correspondence Principle

Before proceeding to show the desired way of uniting the representations, an elaboration of the principle proposed here, called as the *Correspondence Principle* is noteworthy. In order to achieve a unified system of representation, one needs to establish an equivalence relation between (a) PSG and CG and also between (b) DG and CG. The CG derivations would piggyback on PSG constituents in the analyses as the CG derivations proceed as per the constituency relations in PSG with the *wrapping* [1] operation allowed for the functor-argument distance over more than one (constituent) expression, but the CG relations defined on the relevant constituents have to be mapped onto the dependency relations. This warrants a principle that can help unify the DG and CG representations. Therefore, *the correspondence principle* has been proposed.

5 Motivations for The Correspondence Principle

PSG trees fail to capture discontinuous constituents unless the trees are tangled, that is, the *no-crossing constraint* and the *single mother condition* are relaxed as proposed by McCawley (1982). The 'parallel merge approach' too relaxes the *single mother condition*. This relaxation may seem theoretically gratuitous, because this indicates that the *no-crossing constraint* and the *single mother condition* need to be adhered to in cases of non-discontinuity and these very conditions need to be relaxed in cases of discontinuity. This in turn gives rise to two different and separate structural representations for continuous and discontinuous constituents, the former without the above mentioned two conditions relaxed and the latter, with the conditions relaxed, be it PSG trees or tangled trees. On the other hand, DG captures cases of discontinuity, with CG remaining in-between. Given this situation, there arises incompatibility between PSG, CG and DG in analyses of linguistic structures. The *Correspondence Principle* is the 'glue' that can bind the principles of PSG, CG and DG together in a non-superfluous manner for both continuous and discontinuous structures. Once the Correspondence Principle is applied, the need for the *no-crossing constraint* and the *single mother condition* disappears, precisely because all cases demanding these conditions are re-interpreted and re-analysed in terms of functioning of the basic principles of DG and CG united together. Accordingly, this principle would be used for the DG → CG and CG → DG derivations illustrated in the fifth section of this paper.

The Correspondence Principle:

$$A(B^*) \vee A(^*B) \equiv A|B$$

For any two words A and B, A(B*) indicates B is dependent on A and B is to the left of A and A(*B) indicates B is dependent on A and B is to the right of A. Here, '∨' is the logical disjunction, '≡' is a special equivalence sign and A|B indicates that either A or B can be the functor in categorial relations, with '|' indicating the neutral direction of the functor. This implies that the other element will be the argument. The logical relation is that of an *implication*, but not of *an entailment*, because when one element (A or B) is the functor, nothing is said about the other element. In cases where there is a direct dependency relation between the functor and the argument, A and B on the Left-Hand Side (LHS) and Right-Hand Side (RHS) turn out to be the same. However, this is not the case always. In exceptional cases, only either A or B

[1] Wrapping rules usually infix, by way of a sort of swapping, a discontinuous string element in a place where another element was initially located (see for details, Steedman, 1985: 527).

tends to be the same on LHS and RHS, and the other category can vary across sides. If, for example, we suppose that A is the same on both sides, the exact value of B may differ on the LHS and the RHS (that is, B can take a word X, for example, on the LHS, while it takes a word Y, for example, on the RHS). Given this understanding of the *Correspondence Principle*, we shall now turn to the derivations to apply this principle and arrive at a unified system of representation for a Croatian sentence.

6 Towards a Unified Representation: An Illustrative Case of a Croatian Sentence

This section provides an illustration of the unified system of representation for a discontinuous Croatian sentence. An outline of the strategy followed is given below. The following four steps (not necessarily in the same order) are essential: (a) PSG to CG derivation (b) DG to CG derivation (c) CG to DG derivation (d) CG to PSG derivation. For the PSG to CG derivation, the starting point would be the PSG tree, hence the CG derivation is depicted in the PSG tree. For the CG to PSG derivation, each step of the CG derivation is mapped onto an appropriate PSG constituent. The final PSG tree can be derived after the last step of the CG derivation. For the CG to DG derivation, the CG derivation would be the starting point. For a functor-argument relation in each step of the CG derivation the corresponding head-dependent relation is established. The DG tree of the expression can then be drawn based on the outputs of the individual steps. This is where the proposed *correspondence principle* will come into picture. Similarly, for the DG to CG derivation, for each head-dependent relation in the DG graph a functor-argument relation is derived by using *the correspondence principle*. These account for the forward and converse derivations for establishing the equivalence relations CG ≡ PSG and DG ≡ CG. This has been illustrated for the Croatian sentence below:

ii. *Naša je učionica udobna.*
Our is classroom comfortable
'Our classroom is comfortable.' (Van Valin, 2001:88)

In this sentence, discontinuity arises since *je* and *udobna* are not contiguous in the linear order of the sentence.

(iia) A CG derivation in the phrase structure tree (PSG → CG)

Figure 1 depicts the CG derivation of (ii) in its PSG tree and Figure 2 depicts the CG derivation of (ii).

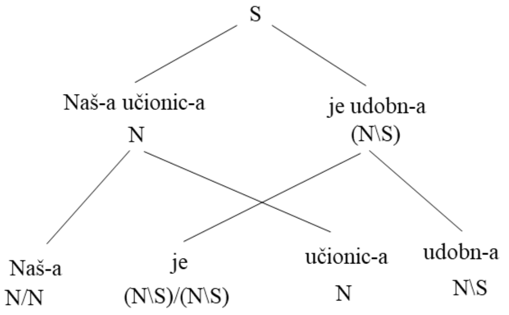


Figure 1. The CG derivation in PSG tree

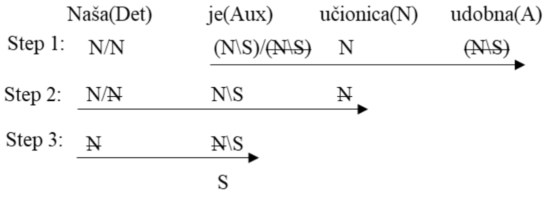


Figure 2. The CG derivation of (ii)

The illustration of Fig 2:

- In step1, the category of *je* is cancelled out with respect to the category of *udobna*.
- In step 2, the category of *Naša* is cancelled out with respect to the category of *učionica*.
- In step 3, the output of step 2 (category of *Naša učionica*) now becomes the input and it is cancelled out with respect to the category of *je udobna*, resulting in the final output S.

It may be noted that though standard PSG trees do not allow for criss-crossing lines in the tree diagrams, the crossing lines are drawn in order to explain how the cancellation of categorial functions can be implemented with the help of the PSG tree as seen in Figure 1. The cancellation of arguments of a function proceeds in accordance with the constituency relations in PSG, as seen in Figure 2. The exact manner in which tree branches are or can be tangled reflects the way categorial derivations can work, thus uniting CG derivations with PSG. That the crossing lines are made insignificant in CG is substantiated by the specification of the series of steps for the categorial derivation of the sentence which is shown right in Figure 2.

(iib) Dependency functions in terms of CG formulae (DG → CG)

The dependency graph for the sentence (ii) is depicted in Figure 3. It may be observed that δ is a dependency valuation function that takes a node as an input and returns a real value as an output (see Levelt, 2008: III:51). If A~B (meaning that A is dependent on B), then δ(A)>δ(B). The real value is set to 0 at the top of the tree, but we can start from 1 at the top of the tree. This function will be useful for recoding CG functor-argument relations in terms of dependency relations as discussed below:

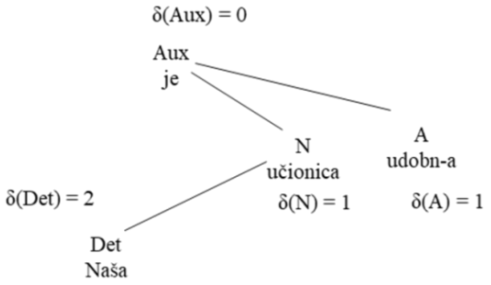


Figure 3. A dependency graph of (ii)

The above figure illustrates the following dependencies:

i. *udobna* is dependent on *je*.
ii. *Naša* is dependent on *učionica*.
iii. *učionica* is dependent on *je*.

Based on Figures 2 and 3, the dependency functions capturing all the functor-argument relations/cancellations can be formulated as follows:

Step 1: δ(Aux)/δ~~(A)~~ ~~δ(A)~~
This step captures the functor-argument relation between *udobna* (A) and *je* (Aux) corresponding to step 1 of the CG derivation in Figure 2. This step builds the meaning conveyed through *is comfortable*.
Step 2: δ(Det)/~~δ(N)~~ ~~δ(N)~~
This step captures the functor-argument relation between *Naša* and *učionica* corresponding to step 2 of the CG derivation in Figure 2. This step builds the meaning conveyed through *our classroom*.

Step 3: δ(Aux)/~~δ(N)~~ ~~δ(N)~~
This step captures the functor-argument relation between the outputs of Step1 and Step2 of the CG derivation. By using the correspondence principle, we have that je(*učionica) ≡ Naša\je. In other words, Aux(*N) ≡ Det\Aux. We can also express it as: A(*B) ≡ B\A (A = *je*). Since 'Naša' and 'je' do not participate in any (direct) dependency relation as seen in Figure 3, the functor-argument relation is constructed through Aux and N in step 3. This step builds the meaning conveyed through *our classroom is comfortable*.

(iic) CG → DG derivation

Here *the correspondence principle* is used to show how the dependency relations can be derived from the categories assigned to the words and the subsequent CG derivation. When each step of the CG derivation is taken into account and expressed in terms of head-dependent relations, the corresponding dependency relation between the functor and the argument can be established.

In this derivation, the equivalence relation is established from RHS to LHS. Hence the CG derivation would be the starting point. The aim is to establish a DG relation for each step of the CG derivation. In other words, for every functor-argument relation, the equivalent head-dependent relation is to be established. Finally, by considering all the head-dependent relations that are established from the CG derivation and other possible head-dependent relations (if any), the DG graph of the sentence can be drawn.

Step 1: The CG relation between 'je' and 'udobna'
In this CG relation, 'je' (Aux) is the functor and 'udobna' (Adj) is the argument. The direction of the argument is to the right. If we consider 'je' (Aux) to be A and 'udobna' (Adj) to be B, the RHS would be Aux/Adj or je/udobna or A/B. There is a direct dependency relation between the functor and the argument - 'je' and 'udobna', with 'je' as the head and 'udobna' as its dependent. Accordingly, the LHS would be je(*udobna) or Aux(*Adj) or A(*B). Thus, the equivalence relation for this CG relation would be:

je(*udobna) ≡ je/udobna. It can also be expressed as: Aux(*Adj) ≡ Aux/Adj or A(*B) ≡ A/B.

Step 2: The CG relation between 'Naša' and 'učionica'

In this CG relation, 'Naša' (Det) is the functor and 'učionica' (N) is the argument. The direction of the argument is to the right. If we consider 'učionica' (N) to be A and 'Naša' (Det) to be B, the RHS would be Det/N or Naša/učionica or B/A. There is a direct dependency relation between the functor and the argument - 'Naša' and 'učionica', with 'učionica' as the head and 'Naša' as its dependent. Accordingly, the LHS would be the following: učionica(Naša*) or N(Det*) or A(B*). Thus the equivalence relation for this CG relation would be: učionica (Naša*) ≡ Naša/učionica. It can also be expressed as: N(Det*) ≡ Det/N or A(B*) ≡ B/A.

Step 3: The CG relation between 'Naša' and 'je'

In this CG relation, 'je' (Aux) is the functor and 'Naša' (Det) is the argument. The direction of the argument is to the left. If we consider 'je' (Aux) to be A, the RHS would be Det\Aux or Naša\je or B\A (B = Det). However, there is no direct dependency relation between the functor and the argument - 'je' and 'Naša'. Rather, 'učionica' (N) is dependent on 'je' (Aux). In other words, Aux(*N) or je(*učionica) or A(*B) [B = 'učionica'] would be the LHS. Since the functor and the argument do not participate in a direct head and dependent relationship, considering 'je' to be A on the RHS would implicitly indicate that B on the RHS is its argument 'Naša' and B on the LHS is its dependent 'učionica'. Thus, the equivalence relation for this CG relation would be: je(*učionica) ≡ Naša\je. It can also be expressed as, Aux(*N) ≡ Det\Aux or A(*B) ≡ B\A. This clearly shows that 'je' (Aux) is the head of 'učionica' (N) but is the functor of the argument 'Naša' (Det).

Thus, combining the DG relations of all the steps, we get the following.

(i) Aux(*Adj) or 'udobna' is dependent on 'je'

(ii) N(Det*) or 'Naša' is dependent on 'učionica'

(iii) Aux(*N) or 'učionica' is dependent on 'je'

Based on these dependency relations, we arrive at the DG graph with the same dependency functions, that is, Figure 3.

(iid) CG → PSG derivation

In this Croatian sentence, discontinuity arises because (i) 'Naša' and 'učionica' and (ii) 'je' and 'udobna' are not contiguous in the linear order of the sentence. However, in a PSG tree, the cancellation of arguments proceeds as per the constituent structure.

Step1 of the CG derivation would mean that Aux and AP form a single constituent AuxP with Aux as the head. The analysis of 'udobna' as a separate constituent 'AP' is drawn from the fundamental principles of PSG as seen in Figure 4.

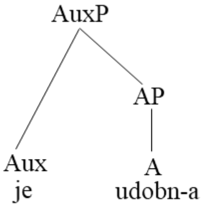


Figure 4. The PSG tree corresponding to step 1 of the CG derivation

Step2 of the CG derivation would mean that Det and N form a single constituent NP with N as the head as seen in Figure 5.

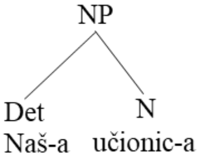


Figure 5. The PSG tree corresponding to step 2 of the CG derivation

Step3 of the CG derivation indicates that NP and AuxP form a single constituent S. Hence, in this step the remaining arguments are cancelled out resulting in the final output S. The corresponding tangled diagram of the sentence is shown below in Figure 6. Accordingly, the corresponding PSG rules are: (i) S → NP AuxP (ii) NP → Det N (iii) AuxP → Aux AP (iv) AP → A.

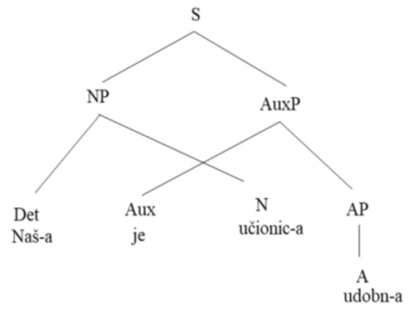


Figure 6. The final PSG tree

(iie) A unified representation (from DG to CG to PSG)

Finally, Figure 7 depicts a unified representation taking into account the constituency relations, dependency relations and the functor-argument relations in the discontinuous Croatian sentence.

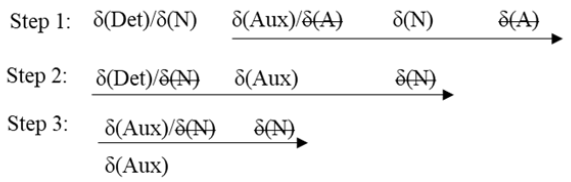

Step 1: δ(Det)/δ(N) δ(Aux)/~~δ(A)~~ δ(N) ~~δ(A)~~
Step 2: δ(Det)/~~δ(N)~~ δ(Aux) ~~δ(N)~~
Step 3: δ(Aux)/~~δ(N)~~ ~~δ(N)~~
δ(Aux)

Fig 7. A unified representation of the discontinuous sentence

In all, the derivations formulated for the above illustration show how the conversions, namely PSG→CG, DG→CG, CG→DG and CG→PSG (not necessarily always in that order), can establish the desired equivalence of representations in those formalisms. Therefore, establishing PSG→CG, DG→CG, CG→DG and CG→PSG is tantamount to establishing PSG≡CG≡DG in their representational principles for natural language constructions.

7 Implications and Conclusion

This paper is an attempt to show how a flexible account of functor-argument relations can be decoded from the rigid constituents of phrase structures and how in turn these functor-argument (categorial) relations can also be formulated in terms of dependency relations. Hence the PSG rules in trees could be redrawn in terms of the CG formula which in turn could be rewritten in terms of the DG functions. This can have far-reaching implications for theories of natural language since most current linguistic theories do adopt and subscribe to constituency-based analyses, although specific treatments of particular phenomena such as labelling phrases may differ. But one emerging conclusion is that not all aspects of natural language (especially syntax) can be accounted for by binary branching and headed rules (see Müller, 2013). The unified system of representation for continuous and discontinuous structures cuts across and in fact (somewhat) neutralizes the traditional distinction between derivational theories (as in Chomsky, 1995) and constraint-based formalisms because both types of formalisms have to define their derivations or constraints on the structural organization of linguistic structures.

Though there have been many solutions proposed so far, to account for discontinuity, we argue that considering an alternative approach will merely add to the solutions existing in the literature. The present unified system of representation differs from the solutions proposed so far in that it attempts to face up to the problem of discontinuity by enabling a direct analysis of discontinuous structures from the basic underlying assumptions in each of the grammar formalisms without introducing any extra assumptions/rules or even constraints. There is no expansion/manipulation of the features of the system as in the case of most of the proposed solutions. This reduces the number of types of structures and strikes a balance between rigidity and flexibility to account for continuous constituents as well as discontinuous constituents which are grammatical. This can comprehensively capture and help analyse both continuous and discontinuous expressions for a range of natural language phenomena including movement/displacement, long-distance dependencies etc. Regarding long distance dependencies, gaps occur when a phrase is fronted. This missing phrase is anaphorically interpreted as in the case of *Wh*-questions, topicalisation etc. The unified system of representation works in a non-local way and, therefore can account for the gaps created as a result of fronting of constituents in long-distance dependencies. This thereby eliminates the need to separately represent the movement of constituents/gaps using arrows or bars. However, an illustration of this is beyond the scope and space requirements of the current paper.

Most importantly, what makes this approach different from the earlier ones is that the equivalence relation between PSG, DG and CG is in terms of the representational descriptions of natural language constructions. Only the representational principles are unified (constituency relations in PSG, head-dependent relations in DG and functor-argument relations in CG), *not* the grammar formalisms as such, in their descriptions of natural language constructions. The very flexibility that PSG can have in allowing for both normal trees and tangled trees (by relaxing the 'no-crossing' constraint) is nothing other than the flexibility DG or CG inherently permits. DG or CG is inherently neutral with respect to line crossing or no-line crossing. Hence the desired flexibility in PSG for continuous and discontinuous structures is an expression or instantiation of the principles of DG/CG itself. Thus, the apparent tension between the three grammar formalisms can perhaps be neutralized by this way of working towards mutually unifying *the most basic principles* of the three formalisms. The novelty of this unified representation is it can help draw correspondences between representations in parsing systems based on head-dependent relations and the parsing systems based on PSG and/or CG relations. Thus, systems of dependency parsing and PSG-constituency-based parsing can have representational interrelation that can help achieve representational economy. A parser can then (de)code dependency parses into constituency parses and vice versa, without any extra burden on computational resources, since one single representational system may suffice. Hence exploring the full range of practical applications of the unified representation is beyond the scope of the current study and would be left as a follow-up.

Another argument substantiating the unified system of representation pertains to the representation of language in our cognitive system. In the real world a speaker of a language with a predominantly continuous system can also learn, comprehend, speak a language with a non-continuous system. Though linguists advance each of these grammatical formalisms on distinct grounds having varying theoretical motivations, it is more likely that for a speaker of any language there exists just one representation in their cognitive system which is equipped to deal with the features of both kinds of systems. However, this is too left open for further study.

References

Barbara Citko. 2011. Symmetry in syntax: Merge, move and labels. Cambridge: Cambridge University Press.

Cathryn Donohue and Ivan A. Sag. 1999. Domains in Warlpiri. In Sixth International Conference on HPSG–Abstracts. 04–06 August 1999, 101–106.University of Edinburgh.

David G. Hays. 1964. Dependency theory: A formalism and some observations. Language, 40: 511–25.

David R. Dowty. 1996. Towards a minimalist theory of syntactic structure. In Bunt, H. & van Horck, A. (Eds.). *Discontinuous constituency*, 11–62. Berlin: Mouton de Gruyter.

Frederick J. Newmeyer. 2001. Grammatical functions, thematic roles, and phrase structure: their underlying disunity. In Davies, W. D. & Dubinsky, S. (eds.). Objects and other subjects, 53-75. Springer, Dordrecht.

Gerald Gazdar. 1983. Phrase structure grammar. Kulas, J., Fetzer, J. H. & Rankin, T. L. (eds.). Philosophy, language and artificial intelligence, 163-218. Dordrecht: Springer.

Guy Barry and Martin J Pickering. 1990. Dependency and constituency in categorial grammar. In Barry, G. & Morrill, G. (eds.), Edinburgh working papers in cognitive science, vol. 5. Studies in categorial grammar, 23-45. Edinburgh: Centre for Cognitive Science, University of Edinburgh.

Guy Barry and Martin J. Pickering. 1993. Dependency and categorial grammar and coordination. Linguistics, 31(5): 855 - 902.

Haim Gaifman. 1965. Dependency systems and phrase-structure systems. Information and Control, 8(3): 304-337.

James D. McCawley. 1982. Parentheticals and discontinuous constituent structure. Linguistic Inquiry, 13(1): 91-106.

James P. Blevins. 1990. Syntactic complexity: Evidence for discontinuity and multidomination. Ph.D.

dissertation, University of Massachusetts at Amherst.

Joachim Lambek. 1958. The mathematics of sentence structure. American Mathematical Monthly, 65(3): 154–170.

Kazimierz Ajdukiewicz. 1935. Die syntaktische konnexität. *Studia Philosophica.* 1, 1-27, transl. *Syntactic connexion* in S. McCall, Polish Logic. Oxford 1967, 207–231.

Ken Hale. 1982. Preliminary remarks on configurationality. In Pustejovsky, J. & Sells, P. (eds.). North East Linguistic Society, 12: 86-96.

Ken Hale.1983. Warlpiri and the grammar of non-configurational languages. Natural Language and Linguistic Theory, 1(1): 5-47.

Kunihiro Iwakura. 1988. Review article on Huck, G. and Ojeda, A. (eds.) 1987. English Linguistics, 5: 329-46.

Lucien Tesniére. 1959. Eléments de syntaxe structurale. Paris: Ed. Klincksieck.

Marie-Catherine de Marneffe and Joakim Nivre. 2019. Dependency grammar. Annual Review of Linguistics. 5: 197-218.

Mark Steedman. 1985. Dependency and coordination in the grammar of Dutch and English. Language, 61(3): 523-568.

Mark Steedman. 1992. Categorial Grammar. Lingua, 90: 221-258.

Mark Steedman. 2014. Categorial Grammar. In Carnie, A., Sato, Y. & Siddiqi, D. (eds.), Routledge handbook of syntax, 670–701. New York: Routledge.

Noam Chomsky. 1957. Syntactic structures. Hague: Mouton de Gruyter.

Noam Chomsky. 1995. The minimalist program. Cambridge: MIT Press.

Peter Austin and Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. Natural Language & Linguistic Theory, 14(2): 215-268.

Rachel Nordlinger. 2014. Constituency and grammatical relations in Australian languages. Koch, H. & Nordlinger, R (eds.). The languages and linguistics of Australia: A comprehensive guide, 215-261. Berlin: Mouton de Gruyter.

R. M. W Dixon. 1972. The Dyirbal language of north Queensland. Cambridge: Cambridge University Press. Dixon, R. M. W. 1977. A grammar of Yidiny. Cambridge: Cambridge University Press.

Robert Van Valin Jr. 2001. An introduction to syntax. Cambridge: Cambridge University Press.

Stefan Müller. 2002. Complex predicates: Verbal complexes, resultative constructions, and particle verbs in German. Stanford, CA: CSLI Publications.

Stefan Müller. 2013. Unifying everything: Some remarks on Simpler Syntax, Construction Grammar, Minimalism and HPSG. Language, 89(4): 920–950.

Willem J. M. Levelt. 2008. Formal grammars in linguistics and psycholinguistics. Vol.3. Amsterdam: John Benjamins Publishing Company.

Yehoshua Bar-Hillel. 1953. A quasi-arithmetical notation for syntactic description. Language, 29: 47–58.

Yehuda N. Falk. 2001. Lexical functional grammar: An introduction to parallel constraint-based syntax. Stanford: CSLI Publications.

Analysis of Well-being in Some Visayan Languages

Randolph G. Catungal
Lagro High School, Q.C.
catungal_randolph@yahoo.com

Abstract

This study is unique because it focuses on language consultants' perceptions of their life satisfaction. This paper looks into their well-being and how it reflects their culture. The questions addressed were, "Are you happy with your life right now?" and "How can you ensure yours and your children's well-being?" After the in-depth interview, three domains have been identified: livelihood, money, and wishes. They are satisfied if they are able to work for a living and mind if their family eats twice or thrice a day. They hope that whatever they failed to achieve in life would be carried on by their descendants. Reduplication, affixation, transfer emphasis, phoneme deletion, and borrowing are analyzed on how some terms related to well-being are being formed in Romblomanon, Waray-Leyte, and Visaya-Mindanao language consultants.

1 Introduction

The experience of pleasant emotions such as happiness and contentment, as well as the development of one's potential, gaining some control over one's life, gaining a sense of purpose, and enjoying meaningful connection, was classified as well-being (Huppert, 2009). According to the World Health Organization (WHO, 2001), well-being is a long-term state that allows a person or a group to grow and develop. It is associated with professional, personal, and interpersonal success, with well-organized people demonstrating higher work productivity, more effective learning, and enhanced creativity. There are ethnolinguistic groups that have the same pronunciation, sound, and spelling but also have different lexicons (words) but with the same semantics (meanings).

2 Methodology

This study utilized a comparative approach to compare the language consultants' perspectives on their well-being at present. It also aims to compare their views of well-being on Romblomanon, Waray-Leyte, and Visaya-Mindanao.

Life satisfaction as one of the measures of well-being is viewed as happiness. This is an initial study because the data collected is limited. The consultants are twelve native speakers of Romblomanon, Waray-Leyte and Visaya-Mindanao aged 25-50. The data gathered were presented and analyzed through componential analysis.

3 Results and Discussions

The terms 'maginhawa', 'mas maginhawa' and 'pinakamaginhawa' are expressed same in Romblomanon, Visaya-Mindanao, and Tagalog. In Romblomanon, well-being or being well is having an adequate occupation or farm, a happy family, and having a good connection towards various types of trade. The life of the Waray-Leyte people is comfortable without bad weather because they make a living from rice and copra. In Waray-Leyte, 'maupay' is the equivalent of the Tagalog words 'ginhawa' and 'maginhawa'. Therefore, having a comfortable life for Waray-Leyte is having a luxurious life. 'Ungod' which is equivalent to 'pinaka' in Tagalog can be associated with 'ubod' which means 'sobra' like 'sobrang ginhawa'. In fact, the different language varieties of Visaya are influenced by different languages in the Philippines such as Tagalog and English. Being comfortable in the life of the Visayan-Mindanao is having a more uplifting life for other people, more uplifting in life. In those days, the datu in Visaya-Mindanao were considered wealthy people in their area.

Table 1 shows the different lexicons for 'ginhawa', 'maginhawa', 'mas maginhawa', and 'pinakamaginhawa' based on the language consultants.

Table 1: Lexicons for Comparing Well-being in Visaya

Visayan Language	Well-being			
	Ginhawa	Maginhawa	Mas Maginhawa	Pinakamaginhawa
Romblomanon	Ginhawa	Maginhawa	Mas maginhawa	Pinakamaginhawa
Waray-Leyte	Maupay	Maupay	Mas maupay	Ungod na maupay
Visaya-Mindanao	Ginhawa	Maginhawa	Mas maginhawa	Pinakamaginhawa

Table 2 shows the consultants'

componential analysis of their livelihood. Farming is the livelihood of the Romblomanons, Waray-Leyte and Visaya-Mindanao. The word "pagsasaka" is used by the Romblomanons and Waray-Leyte for the Tagalog word "pagsasaka". Visaya- Mindanao, on the other hand, uses the word “Pag-ooma” which is equivalent to the Tagalog word “pagsasaka”. "Oma" is the root word with the equivalent of the word "saka" in Tagalog. It uses the so-called process ‘reduplication’ wherein the prefix “pag-“ is added in front of the root word‘saka’ and repeats the phoneme ‘o’ to form a new word just like in ‘pag’ + oma = pag-ooma. This also happens in Tagalog, Romblomanon and Waray-Leyte just like in ‘pag’ + saka = pagsasaka.

Table 2: Componential Analysis of the Visayan Livelihood

Visayan Language	Livelihood (Kabuhayan)	Farming (Pagsasaka)
Romblomanon	Pagsasaka	+
Waray-Leyte	Pagsasaka	+
Visaya-Mindano	Pagooma	+

All the consultants of Romblomanon, Waray-Leyte, and Visaya-Mindanao believed that theycannot achieve their well-being as shown in Table 3 below.

Table 3: Componential Analysis of Over-all Well-being in Visaya

Visayan languge	Over-all well-being
Romblomanon	-
Waray-Leyte	-
Visaya-Mindanao	-

Paz (2008) mentioned that the possible reasons why Philippine languages are similar are due to the cultural influence of the country’s invaders - Spain and the United States. It means theculture has become the foundation of the Filipinos to have one nation, the Philippines.

It is undeniable that the word "ginhawa" has different meanings among the aforementioned varieties of Visaya. See Table 4, below. The ‘maayo’ meaning of ‘ginhawa’ for Romblomanons comes from the word ‘maayos’ which has undergone a morphophonemic change called ‘assimilation’ in which there is a deletion of the phoneme of the original word. The ‘s-’ was removed from ‘maayos’ which is a word in Tagalog so the word ‘maayo’ came out. The equivalent words “Hayahay” and “Paghinga” for the word ‘Ginhawa” in Waray-Leyte and Bisaya-Mindanao are also noted to use such words in Tagalog but “Hayahay” in Tagalog means doing nothing or resting (word denoting action in the present). On the other hand, Waray-Leyte and Visaya Mindanao have the same meaning of ‘may maayos na paghinga’ for the words “Hayahay” and “Paghinga”.

Table 4: Lexicon for Well-being (Ginhawa) in Visaya

Visayan language	Well-being (Visaya)
Romblomanon	Maayo
Waray-Leyte	Hayahay
Visaya-Mindanao	Paghinga

Consultants of the aforementioned Visaya varieties have the same view of ‘money’ as a material object and have the same level of valuation or importance. According to the same articleby Paz (2008), having beautiful houses symbolizes material things in the various ethnolinguistic groups mentioned in the article but not mentioned about the Visayans. See Table 5 below whichshows the call to money as a material object among the Visayans. "Kwarta" is what the Romblomanons, Waray-Leyte, and Visaya-Mindanao call the same word "kwarta". The word "kwarta" is also used in Tagalog but it indicates the influence of the Spaniards on our own language, the Filipino. In fact, money is important to everyone but we are different in valuing it. As a native speaker of Tagalog, I am satisfied that I can buy/pay for my basic needs such as food,water, clothing, electricity, and education. Therefore, the desire for extra money shows luxury and obedience to vice.

Table 5: Componential Analysis of Money (Pera) in Visaya

Visayan Language	Money (Pera)	Valuing (Halaga)
Romblomanon	Kwarta	+
Waray-Leyte	Kwarta	+
Visaya-Mindanao	Kwarta	+

Consultants believed that well-being does not just end with what they currently enjoy prosperity (kaginhawaan/kasaganaan) but well-being is a cycle that must be maintained. Table 6below shows the lexicons comparing their ‘wishes’ (naisin) in Romblomanon, Waray-Leyte and Visaya-Mindanao. Tagalog and Romblomanon use the same word ‘gusto’ as an equivalent to the word nais. Waray-Leyte and Visaya-Romblomanon use the same word ‘pangandoy’ as an equivalent to the word ‘nais’. The equivalent word for 'ninanais' and 'nanaisin' is the same in thenative speakers of Waray-Leyte and Visaya-Mindanao where the root word 'pangandoy' is added by the prefix 'gi-' with the equivalent prefix “ni-” and “na-” in Tagalog. The meaning of “ninanais”and “nanaisin” is the same, which is called ‘gipanganduy’ in Waray-Leyte and Visaya-Mindanao wherein the process of phonemic change is being used. However, romblomanon uses ‘gina-gusto’as an equivalent to the word ‘ninanais’ in Tagalog. They also use the equivalent word ‘gugustuhon’ for the word ‘nanaisin’ in Tagalog. Adds the prefix 'gina-' to the root word 'gusto' to have an equivalent in the Tagalog word 'ninanais' and adds the prefix 'gu-' and the suffix '-hin' to the root word ‘gusto’ and there will also be a shift, a phonemic change in which the emphasischanges when pronounced as in 'gu-' + “gusto” + “-hin-“=gugustohin=gugustohon.

Table 6: Lexicons for Comparing Wishes (Naisin) in Visaya

Visayan Language	Naisin		
	Nais	Ninanais	Nanaisin
Romblomanon	Gusto	Ginagusto	gugustuhon
Waray-Leyte	Pangandoy	Gipanganduy	gipanganduy
Visaya-Mindanao	Pangandoy	Gipanganduy	gipanganduy

Indigenous speakers of Romblomanon, Waray-Leyte, and Visaya-Mindanao also mentioned that they have ‘wishes’ in life not only for themselves but also for their families. Table7 lists the ‘wish’ of the consultants. It can be seen that the consultants of Romblomanon, Waray-Leyte, and Visaya-Mindanao have given different things when it comes to their wishes which increases the expectation of achieving them. The degrees of the adjectives (nais, ninanais, and nanaisin) reflect their wish to achieve them. In Romblomanon, they want to graduate, they wantto have a permanent job and they want to have their own house. This means they have to finishschool first so that they can have their own home in the future. Waray-Leyte and Visaya- Mindanao, on the other hand, have the same wish for life and this is a simple life as well as whatthey wish is abundant life (prosperity). A prosperous life for Waray-Leyte and Visaya-Mindanao consultants means having “maupay” or maayos (good) health.

Table 7: Wishes (Naisin) of Native speakers of Some Varieties of Visaya to their Families

Visayan language	Naisin		
	Nais	Ninanais	Nanaisin
Romblomanon	makapagtapos ng pag-aaral	magkaroon ng permanenting trabaho	magkaroon ng sarilingbahay
Waray-Leyte	simpleng buhay	Masaganang buhay	Masaganang buhay
Visaya-Mindanao	simpleng buhay	Masaganang buhay	masaganang buhay

Paz (2008) mentioned that the Waray ethnolinguistic group has a wish to “make life easier”. This can be seen up to this day. There is probably no one who does not want life to be easier. It shows that consultants are not satisfied with their lives and even in their present lives, as shown in Table 8.

Table 8: Componential Analysis of Life Satisfaction

Visayan language	Life Satisfaction
Romblomanon	-
Waray-Leyte	-
Visaya-Mindanao	-

Paz (2008) noted in the same article that many ethnolinguistic groups believe in the worldof spirits. In fact, he mentioned performing rituals and chants as a means to drive them away from pain and attain peace, contentment, and order. For the purpose of this research paper, the well-being of the consultants was looked at because they believe that when you live well you have also achieved peace of your heart and your contentment.

Table 9 shows the consultants’ beliefs about achieving well-being. Romblomanon consultants need to study hard to achieve thedesired well-being. For Waray-

Leyte consultants, in order to achieve prosperity, they need to work hard to provide for the daily needs of their families as well as the proper care of children because they believe that children, so as not to grow arrogant, depend on the proper nurturing of them. Therefore, in order for Waray-Leyte and Visaya-Mindanao consultants to provide for their families, they need to work well.

Table 9: Ways to achieve Well-being

Visayan language	Well-being strategy
Romblomanon	Pag-aaral ng mabuti
Waray-Leyte	Paghahanapbuhay at Wastong Pag-aaruga sa mga anak
Visaya-Mindanao	Paghahanapbuhay

The consultants of Romblomanon and Visaya-Mindanao have their own description of the well-being of the children and this is through their wishes. In the question: how can they achieve well-being? This ethnolinguistic group provided answers and according to them, their children will continue what was left to them. The Waray-Leyte consultant, on the other hand, believes that while their children are still on them, it is their obligation to raise them. There are also consultants of the same ethnolinguistic group who say that if their children, would not graduate, they will continue it as mentioned in the consultants of Romblomanon and Visaya- Mindanao.

Table 10 provides the componential analysis of the children's well-being of the aforementioned consultants.

Table 10: Componential Analysis of Well-being of Children

Visayan language	Children's well-being
Romblomanon	-
Waray-Leyte	+/-
Visaya-Mindanao	-

Waray-Leyte and Visaya-Mindanao consultants believe in a superstition that their farmingwill be more efficient (prosperous) if they perform a ceremony (ritual) they call offering. The offering is the offering of the body of the slaughtered pig/chicken and its blood and is performedat six o'clock in the afternoon. There are two people involved, the owner (mag-ooma) of the omahan (farmland) and the prayer who is usually also a 'mag-ooma'. Nowadays, chickens are often slaughtered because pork is so expensive. Back then the pig was slaughtered because it was the custom of the Ancestors (ancestors). The slaughtered pork/chicken is placed in the container (bilao) of the harvested rice and its blood is dripped in each corner of the rice field and then the bilao is placed in the middle of the house where the slaughtered pigs/chickens are placed. And this will be followed by prayer. The prayer includes their wishes in the prosperity spirit that their homestead grows well as well as when the harvest comes. After the ritual, the slaughtered pork/chicken will be cooked and shared by those in the household. In the same article, Paz (2008) mentioned the spirit world as a separate domain of well-being but in the present paper, the spirit world is part of their claim in relation to their livelihood, farming. Nevertheless, Romblomanons do not believe in Spirits in order to have prosperity. According to them, the spirit you often only hear in other people's stories, street stories, and superstition. If you also look at it, perhaps the spirit and superstition are connected because in Romblon, following superstitions brings misfortune that is said to be brought by evil spirits. For them, it is man himself who makes his luck and misfortune. Sometimes, there are things we don't get and we consider bad luck but it's just not really for us and not at the right time.

Table 11 shows Prosperity Spirit's componential analysis of the aforementioned ethnolinguistic groups.

Table 11: Componential Analysis of Prosperity/Spirit In Visaya

Visayan language	Prosperity/Spirit
Romblomanon	-
Waray-Leyte	+
Visaya-Mindanao	+

The well-being of Romblomanon, Waray-Leyte, and Visaya-Mindanao speakers is determined by their source of income (livelihood), money, and demand (wishes). As a Tagalog speaker, I see work as the most important factor in having a good life because my life satisfaction is also dependent on it (pleasure), but pleasure is not permanent. If my job is good, my money will be fine, as will my life's demands. Table 12 shows the componential analysis of the well-being of the Romblomanon, Waray-Leyte, and Bisaya-Mindanao consultants. These consultants believe they are not yet experiencing happiness. Huppert's (2009) article supports the notion that well-being is more than just being happy and satisfied with life, as does the World Health Organization's article that well-being is a permanent state that allows a person or group of people to grow and develop.

Table 12: Componential Analysis of Specific Well-being in Visaya

Visayan language	Specific Well-being		
	Kabuhayan (livelihood)	Pera (money)	Naisin (wishes)
Romblomanon	-	-	-
Waray-Leyte	-	-	-
Visaya-Mindanao	-	-	-

4 Conclusions

Most of the ethnolinguistic groups included in this study are surviving through agriculture. For consultants, a good harvest determines well-being. This is supported by the same article by Paz (2008).

Therefore, the well-being of the language consultants of Romblomanon, Waray-Leyte and Visaya-Mindanao is based on their livelihood (hanapbuhay), money (kwarta) and wishes (kahilingan/naisin).

As a native speaker of Tagalog, I also consider occupation as a primary basis of a good life because it depends on my life satisfaction (ginhawa) but satisfaction is not permanent. If I have a good job, then I will get a good income so I can meet my expectations.

Acknowledgments

I am grateful to all the language consultants who participated in my research, especially my mentor.

References

Huppert FA. (2009). Psychological well-being: evidence regarding its causes and consequences. Appl Psychol Health Well Being;1(2):137–64. https://doi.org/10.1111/j.1758-0854.2009.01008.

Paz, C.J. (2008). Ginhawa, kapalaran, dalamhati: Essays on well-being, opportunity/destiny, andanguish. Quezon City: The University of the Philippines Press

World Health Organization. The world health report 2001: mental health: new understanding, new hope.Geneva: World Health Organization; 2001

Appendix A. Morphophonemic Changes in Some Lexicons Related to Well-being in Some Visayan Languages

VISAYAN LANGUAGE	Root word	Prefix	Suf-fix	New word	Morphophonemic Change
Romblomanon	saka ayos gusto	pag- n/a gina- gu-	n/a n/a n/a -hin	pagsasaka maayo ginagusto gugustohon	Reduplication Phoneme deletion Affixation Transfer emphasis
Waray-Leyte	saka hayahay pangandoy	pag -n/a gi-	n/a n/a n/a	Pagsasaka Hayahay gipanganduy	Reduplication Borrowing Transfer emphasis
Visaya-Mindanao	oma hinga panganduy	pag- pag- gi-	n/a n/a n/a	pag-ooma paghinga gipanganduy	Reduplication Affixation Transfer emphasis

The Information Packaging of the Do-Constructions in Chinese, Russian, and Czech

Chui, Kawai
Department of English
National Chengchi University
kawai@nccu.edu.tw

Yeh, Hsiang-lin
Department of Slavic
Languages and Literatures
National Chengchi University
verayeh@nccu.edu.tw

Lin, Melissa Shih-hui
Department of Slavic
Languages and Literatures
National Chengchi University
shihhui@nccu.edu.tw

Abstract

This study investigated the *do*-constructions in Chinese, Russian, and Czech, a predicate-argument structure comprised of the light verb ‘to do’ - *zuò* in Chinese, *delat'* in Russian, and *dělat* in Czech - and a verbal noun as the head in the accusative role, considering the linguistic traits and pragmatic use of the constructions in spoken and written discourse. The corpus results attested that the three languages not only have lexical and grammatical equivalences, they also demonstrate a functional equivalence in packaging information to define a type of action within the construction. Similar lexico-grammatical strategies are employed to encode tense and aspectual information of the predicates and various kinds of information about the nominal heads. The preference of the *do*-usage in the written genre is unequivocal in Chinese and Russian, suggesting that the structural change could have started as a writing style. The relative novelty of the *do*-usage to communicate generic or specific action events in Czech is evidence of language-specificity in pragmatic use.

1 Introduction

Light verb constructions, such as ‘to take a turn’, ‘to give a rating’, ‘to make a second attempt’, and ‘to do a quick change’, are cross-linguistic structures which have been studied in many languages including American English, British English, Irish English, Malaysian English, Zapotec, Spanish, Italian, Russian, Czech, Lithuania, Urdu, Persian, Japanese, Korean, and Chinese. The head nouns in the accusative position are ‘verbal nouns’ and the constructions as a whole may be equivalent to the use of the head nouns as full-fledged verbs, also called ‘heavy verbs’, such as ‘to turn’ versus ‘to take a turn’, ‘to rate’ versus ‘to give a rating’, ‘to attempt (to do something) for the second time’ versus ‘to make a second attempt’, and ‘to change quickly’ versus ‘to do a quick change’. A variety of issues were discussed from various approaches, among which the constructions were investigated in regard to argument or valency structures (Yim, 2020; Kettnerová, 2021; Kettnerová and Lopatková, 2020; Lin, 2014; Ronan and Schneider, 2017; Tadao, 2000). The corpus-based or descriptive approaches were employed to investigate the lexical, structural, semantic, and pragmatic properties of the light-verb constructions (Cuervo, 2010; Hernández, 2008; Huang and Lin, 2012; Huang et al., 2014; Maiko, 2020; Martínez Linares, 2013; Nolan, 2015; Ong and Rahim, 2021; Radimský, 2010; Ronan, 2014; Kovalevskaite et al, 2020), the linguistic distinction between the light verb and heavy verb usages (Beam de Azcona, 2017; Evteeva, 2017; Lu et al., 2020; Radimský, 2010; Tadao, 2000), the occurrences of the constructions in speaking and writing (Sundquist, 2020), the historical development of the constructions (Buckingham, 2014; Sundquist, 2018; Yim, 2020), the acquisition of the constructions in L2 contexts (Maiko, 2019; Sanromán Vilas, 2019), and the processing of light-verb structure (Wittenberg and Piñango, 2011). Little attention, however, has been paid to how the construction pairs the lexico-grammatical structure with meaning and function (Croft, 2014; Goldberg, 1995). As distinct from the usual predicate-argument combination where the event type is determined by the predicate, it is the nominal argument of the *do*-construction that denotes a type of event in discourse.

This study investigates the light verb 'to do' forming a predicate-argument structure with a verbal noun as the head in the accusative role. This is called '*do*-construction' here. See the underlined parts in these English examples 'I did some swimming and headed home', 'I would do less correcting and more connecting', and 'We obviously need to do a lot of praying' from SKELL (skell_3_10 v1.8). The *do*-constructions in Mandarin Chinese ('Chinese' for short), Russian, and Czech are illustrated below. In Example 1, about psychological simplification, *zuò* 'to do' is the main verb and *jiǎnhuà* 'simplify' is the verbal noun as the head of the direct object which is quantified by *yīxiē* 'some' and characterized by *xīnlǐshàng* 'psychological'. In the Russian Example 2, 'to do author citations' is represented by the main verb *делать* 'to do' and the accusative form of the head noun *ссылки* 'citations' is qualified by *авторов* 'author'. In the Czech *do*-construction in Example 3, the verb *dělat* 'to do' and the accusative head noun *přehled* 'overview' as characterized by *dokonalý* 'perfect' together refer to the act of perfect overview.

(1) *nǐ* *kěnéng* *xūyào* ***zuò*** *yīxiē*
2SG may need do some
xīnlǐshàng *de* ***jiǎnhuà***
psychological DE simplify
'You may need to do some psychological simplification.'

(2) ***Делайте*** *хотя бы* ***ссылки*** *на*
do-IMP at least citation-PL-ACC to
авторов.
authors
'At least do author citations.'

(3) ***Udělali*** *dokonalý* ***přehled****.*
do-PST.PFV.3PL perfect overview
'They did a perfect overview.'

The lexical and structural similarities of the *do*-constructions across Chinese as a Sino-Tibetan language, Russian as an East Slavic language, and Czech as a West Slavic language are not a coincidence. According to Natural Semantic Metalanguage (NSM), there are basic and universal semantic primitives that are conceptually simple and irreducible. "Evidence indicates that this highly constrained vocabulary and grammar has equivalents in all or most languages of the world" (Goddard and Wierzbicka 2014:86). An inventory of semantic primes was proposed by Wierzbicka and colleagues as universal semantic fundamentals which been examined across a wide range of typologically different languages including Arrernte, Chinese, Ewe, French, German, Italian, Japanese, Lao, Malay, Mangaaba-Mbula, Maori, Polish, Russian, Spanish, and Yankunytjatjara (see the details in Goddard and Wierzbicka, 2014). 'DO' is a semantic primitive, and the direct lexical realization of this fundamental and universal concept of action is the *zuò*-verb in Chinese, the *delat'*-verb in Russian, and the *dělat*-verb in Czech. The lexical and grammatical behaviors of the three *do*-words as full-fledged verbs are not only identical, the verbs have also been undergoing a similar grammatical development and *do*-constructions are evolved. What remains obscure are the linguistic nature and the pragmatic use of the evolved structure in spoken and written discourse across the three languages.

The present study takes up the issue and asks how the basic semantic notion of DO engages in developing a widely-used structure. The lexical and structural equivalences of DO in Chinese, Russian, and Czech allow for cross-language investigation of the linguistic properties and the pragmatic use of *do*-constructions by carrying out a corpus analysis of *do*-cases derived from the major spoken and written genres. These research questions are addressed – What are the linguistic properties of the *do*-constructions in Chinese, Russian, and Czech? Are there genre differences between speaking and writing in regard to linguistic traits and occurrence rate? Is there language specificity in the pragmatic use of *do*-constructions? The corpus results enable establishment of a common functional construal of information packaging and discussion of the directionality of the structural spread and historical development of *do*-constructions.

2 The corpora and methods

Language use may vary between speaking and writing. For instance, the 3-word and 4-word lexical bundles predominated in spoken discourse, but a different combination of invariable function words and an intervening content word was prevalent in written academic discourse (Biber, 2009). The collocates for the verbs *have*, *make*, and *take* in conversation were also found distinct from those in informational writing (Conrad and Biber, 2009). The present study thus separates the spoken and written data for analysis. The data are drawn from the Corpus of Contemporary Taiwanese Mandarin 2017 (COCT), the Russian National Corpus, and the Czech National Corpus. First, the COCT documents written data from 1986 to 2017 in the areas of philosophy, religion, science, applied sciences, social sciences, history, geography, language, literature, arts, commerce, and recreation, totaling about 250-million words. The 2007-2014 spoken data consists of 6.6-million words from the sub-titles of Da Ai Journal, a TV program that documents inspiring stories of people and events around the world in areas of law, politics, finance, current events, science, living, fashion, culture, education, and arts. Second, the written data of the Main Corpus of the Russian National Corpus consists of 337-million words collected from fiction and news texts, and the spoken data, totaling 13.3-million words, are the recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies. Data for this study are derived from 1981 to 2019. Last, the written corpus of the Czech National Corpus comprises 4255-million words collected between 1989 and 2014, and the spoken corpus consists of 7-million words produced in informal settings from 2002 to 2017.

The selection of data for this study met the criteria that the predicate of a clausal statement is the *do*-verb, namely *zuò* in Chinese, *делать (delat')* in Russian, and *dělat* in Czech, and the direct object comprises a verbal head noun which can be used as a full-fledged verb in other contexts. For instance, *jiǎnhuà* 'simplify' is the nominal head in Example 1, but the main verb in this statement of *wǒmen jiǎnhuà le jiàokēshū* 'We simplified textbooks.' Russian and Czech show the same usages, in that the accusative form *ссылки* in Example 2 is used as a verb in *Она ссылается на научные исследования* 'She cites scientific research', and the accusative head noun *přehled* 'overview' in Example 3 is a verb in *Z vrcholku hory lze přehlédnout široké okolí* 'From the top of the mountain you can overview the wide surroundings.' Upon this common lexical and grammatical foundation, cross-linguistic results are comparable. The search functions in the corpora were used for data selection. The linguistic analysis of the selected data for each language was carried out by a first analyst and then checked and revised by a second analyst. Table 1 presents the sets of *do*-cases for the study. The Slavic languages consistently have a lower occurrence rate of *do*-cases than Chinese, whether in writing or in speaking. The overall frequencies in Chinese outnumber Russian by 6.6 times and Czech by 19.5 times, and the use of this construction is 3 times more prevalent in Russian than in Czech. The written cases predominate at 76.4% in Chinese and at 85.9% in Russian, whereas Czech shows a close distribution of the data across writing and speaking. These quantitative differences demonstrate that the semantic primitive of DO has been undergoing the same structural development across languages yet not in the same pace.

	Written	Spoken	Total
Chinese	15949	4925	20874
	76.4%	23.6%	100%
Russian	2725	448	3173
	85.9%	14.1%	100%
Czech	548	521	1069
	51.3%	48.7%	100%

Table 1: *Do*-cases in Chinese, Russian, and Czech.

3 The information packaging of *do*-constructions

The *do*-construction consists of two parts. The first part is the *do*-verb which means 'to act'; the second part is the noun phrase in the role of direct object. The nature of the action event is context dependent and determined by the encoding of information within the construction. The *do*-verbs in the Slavic languages are marked with tense and aspect information in all the cases. See the use of *делал* in Russian indicating the past and the imperfective aspect of the act of doing corrections in Example 4, and *udělal* in Czech showing the past perfective action of doing fake recordings in Example 5. Chinese, however, tends to encode these two types of information outside the construction by use of adverbials like *zuótiān* 'yesterday' and *míngtiān* 'tomorrow', *yǐjīng* 'already', *céngjīng* 'once', and *yīzhí* 'continuously'. The occurrence rate of aspect markers within the *do*-construction, like the perfective *le* as in

Example 6 about having done a very bad guide, the experiential *guò*, or the durative *zhe*, is low at 29.9% of the total 20874 cases.

(4) *Он* ***делал*** *бесконечные*
he do-PST-IPFV endless
исправления.
corrections
'He did endless corrections.'

(5) *Proč* ***udělal*** *ty falešné*
why do-PST.PFV.3SG these fake
zápisy do svého deníku?
recordings into own diary
'Why did he do these fake recordings into his own diary?'

(6) *wǒ céngjīng duì háizi de chuàngyì*
1SG once to children DE creativity
zuò le *yī gè hěn bù hǎo de*
do PRF one CL very NEG good DE
yǐndǎo
guide
'I once did a very bad guide to children's creativity.'

In the accusative position of the construction, the verbal noun as the head of direct object functions to represent an action event, and the noun phrase as a whole refers to a generic or specific event in discourse. A generic event refers to a general situation that is encoded by a bare nominal head without semantic characterization in the *do*-construction, such as *zuò chuànzhū* 'to do bead stringing' in Chinese (Example 7), *делать подтяжку* 'to do facelifting' in Russian (Example 8), and *udělali zátah* 'to do pulling' in Czech (Example 9).

Generic events – bare nouns

(7) *nóngfū xiàwǔ máng-wán nóngshì*
farmer afternoon work-finish farming
hòu jiù huì zuò zài liángtíng shàng
after then will sit at pavilion on
zuò chuànzhū
do bead
'The farmer, after finishing farming in the afternoon, would sit in the pavilion and do beading.'

(8) *А ваша жена* ***делала***
and your wife do-PST.IPFV.3SG
подтяжку*?*
facelifting
'Did your wife do a facelifting?'

(9) ***Udělali zátah****, prohledali*
do-PST.PFV.3PL pull search-PST.PFV.3PL
a našli spoustu zásob.
and find-PST.PFV.3PL a lot of stock
'They did a pulling, searched and found a lot of stocks.'

A specific event, on the other hand, refers to a particular situation encoded with nominal qualification. Similar lexico-grammatical strategies, which are broadly categorized into definiteness, quantity, possession, and other qualifying properties, are employed to define specific events in Chinese, Russian, and Czech. See the following examples for the four types of strategies in the languages. First, definite referents of the *do*-events are marked by demonstrative words as in the Chinese 'do these three kinds of recycling' (Example 10), the Russian 'do such kind of recording' (Example 11), and the Czech 'do this discovery' (Example 12). Second, quantified referents are encoded by quantifiers or numerals such as 'do a little improvement' (Example 13), 'do one more stopping off' (Example 14), and 'do one adjustment' (Example 15). Third, the possessive information has to do with someone in possession of the nominal referents as in 'do our planting' (Example 16), 'do his own warnings (Example 17), and 'do my own smiling' (Example 18). Finally, other qualifying properties provide attributive information as in 'do a brief and seemingly meaningful pausing' (Example 19), 'do a witty literature review on nationalism' (Example 20), and 'do a significant smiling' (Example 21).

Specific events – definiteness

(10) *duì wǒ dàgài jīběnshàng huì zuò*
to 1SG probably basically will do
zhè *sān lèi de huíshōu*
this three kind DE recycle
'To me, basically, I probably will do these three kinds of recycling.'

(11) *Будучи на краю гибели ученый*
being on verge death-GEN scientist
делает в своем
do-PRS.IPFV.3SG in one's own
дневнике ***такую*** *запись: <…>.*
diary such recording

‘Being on the verge of death, the scientist does such kind of recording in his diary.’

(12) *Když ona udělala* ***tenhle***
when she do-PST.PFV.3SG this
objev a zavolala mi.
discovery and call-PST.PFV.3SG me
‘When she did this discovery and called me.’

Specific events – quantity

(13) *jiāzhǎng hěn lèyì wèile háizǐ ānquán*
parent very happy for child safety
zuò ***yīdiǎndiǎn*** *gǎishàn*
do a little improve
‘Parents are happy to do a little improvement for the safety of the child.’

(14) *Через полкилометра, на перекрестке –*
after half a kilometer at intersection
направо! Там делаем
to the right there do-PRS.IPFV.1PL
еще ***один*** *заход!*
more one stopping off
‘After half a kilometer, at the intersection – to the right! We're doing one more stopping off there!’

(15) *Určitě tam udělám*
definitely there do-PRS.PFV.1SG
úpravu ***jednu****.*
adjustment one
‘I will definitely do one adjustment there.’

Specific events – possession

(16) *wǒmen shì zài wúchénshì lǐmiàn*
we COP at dust-free room inside
zuò ***wǒmen*** *de zāizhòng*
do our DE planting
‘We do our planting in a dust-free room.’

(17) *А то Министерство здравоохранения*
Otherwise Ministry health-GEN
обязательно делало бы
definitely do-PST.IPFV.3SG would
свои *предупреждения.*
one’s own warning.PL
‘Otherwise, the Ministry of Health would definitely do his own warnings.’

(18) *Udělal jsem* ***svůj*** *bolestný úsměv*
do-PST.PFV.1SG one’s own painful smile
kolem úst.
around mouth
‘I did my own painful smiling around my mouth.’

Specific events – qualifying properties

(19) *tā zuò le yī gè* ***jiǎnduǎn ér shì***
he do PRF one CL brief and seem
yǒuyìhán *de tíngdùn*
meaningful DE pause
‘He did a brief and seemingly meaningful pausing.’

(20) *Джон Бройи делает*
John Breuilly do-PRS.IPFV.3SG
остроумный *обзор* ***литературы***
witty reviewing literature-GEN
о национализме *на глубину в*
about nationalism to depth at
четыре десятилетия.
four decades
‘John Breuilly does a witty literature review on nationalism to the depth of four decades.’

(21) *Udělal jsem* ***významný*** *úsměv.*
do-PST.PFV.1SG significant smile
‘I did a significant smiling.’

Most of the *do*-cases are specific action events, taking up 89% of the total in Chinese, 61.5% in Russian, and 66.4% in Czech. Differences are evident between writing and speaking. First, the frequency distribution of generic cases across the written and spoken data is about equal in Chinese and Czech, while Russian has the large majority of cases in the written texts. Second, the three languages align to show that specific events are the majority in writing, and the mean proportions are much higher in Chinese at 79.9% and Russian at 86.5% than in Czech at 56.3%. See Table 2.

Generic events	Written	Spoken	Total
Chinese	1103	1190	2293
	48.1%	51.9%	100%
Russian	1036	185	1221
	84.8%	15.2%	100%
Czech	148	211	359
	41.2%	58.8%	100%
Specific events			
Chinese	14846	3735	18581
	79.9%	20.1%	100%
Russian	1689	263	1952
	86.5%	13.5%	100%
Czech	400	310	710
	56.3%	43.7%	100%

Table 2: Frequency distribution of generic and specific *do*-cases in spoken and written data.

Considering the types of action events, among the 1453 types of verbal nouns in the written texts and 820 in the spoken texts in Chinese, 470 types of action are found in both writing and speaking. A larger variety of verbal nouns are used in the written mode at 67.7% than in the spoken mode at 42.7%, suggesting the vitality of the *do*-construction in written communication. The two Slavic languages have smaller sets of common types of action, a total of 85 in Russian and 40 in Czech. Like Chinese, Russian includes a lot more diverse types in writing at 70.8% than in speaking at 43%. Czech shows the opposite, in that there is a higher proportion of action types not found in the written texts at 72%. In regard to token frequencies, the shared action types crucially account for the large majority of cases in Chinese and Russian - 83.9% of all the Chinese written data and 86.7% of the spoken data; 80.8% of the Russian written data and 82.2% of the spoken data. In Czech, the highly repetitive use of the common types is seen only in the written cases at 89%. A large portion of the spoken cases, at 49%, demonstrate a much wider variety of action types in speech communication.

4 General discussion

Across Chinese, Russian, and Czech, the occurrence rates of *do*-constructions vary but the form, meaning, and function are equivalent. Croft (2014: 19) noted that "a construction (or any construction) in a language (or any language) used to express a particular combination of semantic structure and information packaging function." The *do*-construction comprises information that expresses a type of action which is denoted by the verbal noun rather than the *do*-verb, and the nature of the event has to do with the packaging of information about the accusative head in the context of use. The tense and aspectual information of the *do*-verb and the various kinds of information about the accusative head nouns together are essential to communicate a type of action event or a specific action event that is of interest in the context of use. This functional construal of the *do*-structure is evident in Chinese, Russian, and Czech, and the encoding strategies are cross-linguistically equivalent.

Across the written and the spoken texts, the structural and distributional analyses attested to the preferred communication of specific action events in the *do*-constructions across languages. In terms of token frequencies, the two Slavic languages are similar to be less productive than Chinese; still, the occurrences across text genres are divergent between the two languages. The prevailing use of this grammatical structure in writing suggests that the encoding strategies have come to be adopted more readily as a writing style in Chinese and Russian. As to Czech, whether the *do*-construction tends to be a writing style or a speaking manner is not clear because of the relatively low occurrence rate. Regarding type frequencies, a verbal head being used in both the written and spoken texts was counted as a type. The Chinese data yielded a total of 470 verbal nominal heads, accounting for 83.9% of the cases in writing and 86.7% in speaking. Similar results are seen in Russian - 80.8% of the cases in writing and 82.2% in speaking refer to a set of 85 action events. Czech has a smaller set of 40 nominal heads that occurred in both types of text. Their occurrences account for 89% of the written data but only 51% of the spoken data. Taken the results together, the types of action that were brought up for discussion in the two types of discourse are considered to be more acceptable by language users and likely function as replicators that propagate the development of the *do*-structure. Synchronically, the cross-genre development is language-specific. The spread of *do*-usages was similar across written and spoken discourse yet only in Chinese and Russian. The Czech language manifests much novelty and diverseness in use of the *do*-construction in spoken communication.

Historically, the use of the *do*-construction in Chinese was far from common before the 20th century. From the Academia Sinica Ancient Chinese Corpus, 296 cases were derived, such as *zuò bùshī* 'to do almsgiving' in *The Water Margin: Outlaws of the Marsh* written in the late 14th century, *zuò gè zhèngjiàn* 'to do witnessing' in *Dream of the Red Chamber* in the 18th century, and *zuò jūtíng* 'to do short-time staying' in *The Scholars* in early 19th century. Further, a large portion of the cases at 48.6% were derived from the texts of *The Water Margin* which is known to be written in vernacular Chinese and considered as close to the spoken language. In a former study of the *do*-structure, the written data in the 20th century drawn from the 11-million-word Academia Sinica Balanced Corpus of Modern Chinese 4.0 (Sinica

Corpus) and the 382-million-word Chinese GigaWord 2 Corpus yielded a total of 3117 cases from 1981 to 2007 (Chui, 2018), and, in this study, five times more data from 1986 to 2017 were retrieved from the Corpus of Contemporary Taiwanese Mandarin. The diachronic data together reveal a contrast in the use of the *do*-construction before and after the 20th century.

In the Slavic languages, the historical Russian data in the Russian National Corpus are available from the 18th and the 19th centuries and a total of 2493 *delat'*-cases were drawn. 86.4% of the data were found in the texts between 1801 and 1900, such as *делать препятствие* 'to do hindering' (year 1775) and *делать великие описания* 'to do great descriptions' (1761-1765). Since the usage did not show a surge in the 20th century, 3173 cases in total, the Russian *delat'*-construction appears to be used earlier than Chinese in the 19th century, such as *делать наблюдения и открытия* 'to do observations and discoveries' (year 1867). The developmental tendency is similar in Czech. From the Czech National Corpus, a total of 134 *dělat*-cases were drawn from 1301 to 1900. Most of the data at 70.1% were from the texts in the 19th century, such as *dělat konec* 'to do an ending' (year 1894), after which the usage spreads gradually. See Table 3. In sum, the diachronic development of this grammatical structure since 1301 in Chinese and Czech supports Feltgen et al.'s (2017) claim that there could be a latency period of a change prior to the expansion of the use in the 19th or 20th centuries. In the present time, an S-curve for the development of the *do*-constructions is not seen due to the lack of a slow tailing off. It is also possible that the S-curve is not universal (Ghanbarnejad et al., 2014).

Chinese	Russian	Czech
1301-1900 N = 296	1701-1900 N = 2493	1301-1900 N = 134
1986-2017 N = 20874	1981-2019 N = 3173	1989-2017 N = 1069

Table 3: *Do*-cases in historical data.

Language change is initiated by language use (Feltgen et al., 2017). In the basic evolutionary model of language change, speakers replicate linguistic structures in utterances while interacting with other speakers, suggesting the usual directional spread from speech to writing (Blythe & Croft, 2012). The rise of the use of *do*-construction over the past 36 years (1981-2017) in Chinese and that of the *delat'*-construction in the past 38 years (1981-2019) in Russian demonstrate the predominant use in the written texts. The contemporary corpus data of these two languages further confirm the directionality of the structural change from writing to speaking as proposed in Chui (2018). In the literature of language change, cross-linguistic evidence was abundant to support the typical path of linguistic development from spoken to written discourse (Biber & Gray, 2011; Bybee & Hopper, 2001; Croft, 2000; Good, 2008; Hruschka et al., 2009). The reverse direction of change for the development of *do*-constructions is not common but by no means impossible. First, Biber & Gray's (2011) study attested that English complex noun phrases started in academic writing, but the structures did not spread to conversation. Second, if the change had started in the speaking environment, the use of the *do*- and the *delat'*-structures should have been more frequent in the spoken data. The statistics in Table 1 show the otherwise that the *do*-cases in writing are three times more frequent than those in speaking in Chinese, and the *delat'*-cases are six time more common in Russian. In Czech, the occurrence rates of the *dělat*-cases are about the same between the two genres, and, due to the relatively small amount of data, whether the language conforms to the common direction of linguistic spread that the spoken language affects the written language remains inconclusive.

Finally, the *do*-bare noun combinations are related to the use of the nominalized form as a full-fledged verb, in that both forms refer to a generic type of action, such as *zuò băguān* 'to do gatekeeping' versus *băguān* 'to gatekeep', *делать наколку* 'to do tattooing' versus *наколоть* 'to tattoo', and *dělat procházku* 'to do a walk' versus *procházet* 'to walk'. They are, however, different construals of the same experience. In the evolutionary framework for language change based on Hull's general analysis of selection for evolutionary systems (Blythe & Croft, 2012; Croft, 2000; Hull 1988, 2001), the canonical interactor in language change is the user who chooses what to say and how to say it. In other words, the *do*-usage and the full-verb usage could be the alternatives at the user's disposal in communication. It remains to be seen whether the two usages engage in linguistic competition and whether the *do*-form bears any social or individual values as distinct from the full-verb usage.

In conclusion, the present study presented corpus evidence that DO not only has lexical and grammatical equivalences in Chinese, Russian, and Czech, this semantic primitive further demonstrates a less-known equivalence in structural change dated back to the fourteenth century in Chinese and Czech and to the eighteenth century in Russian. Along with the change is the evolvement of the pragmatic function of packaging information for defining a type of action within the *do*-constructions by virtue of common lexico-grammatical strategies. The preference of the *do*-usage in the written genre is unequivocal in Chinese and Russian, leading to our conjecture that the structural change could have started as a writing style. The relative novelty of the do-usage to communicate generic or specific action events in Czech is evidence of language-specificity in pragmatic use.

Acknowledgments

This research was supported by the College of Foreign Language and Literature, National Chengchi University, with the grant from the Higher Education SPROUT Project of the Ministry of Education of the Republic of China.

References

Beam de Azcona, Rosemary G. 2017. From the heavy to the light verb: An analysis of tomar ‘to take’. *Lingvisticæ Investigationes*, 40(2):228–273.

Biber, Douglas and Bethany Gray. 2011. Grammar emerging in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15:223–250.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3):275–311.

Blythe, Richard and Croft William. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 88:269–304.

Buckingham, Louisa. 2014. Light verb constructions in the history of English. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière and Lieven Vandelanotte, editors, *Corpus Interrogation and Grammatical Patterns*. John Benjamins, Amsterdam, pages 15–34.

Bybee, Joan and Paul J. Hopper. 2001. *Frequency and the Emergence of Linguistic Structure*. Oxford: John Benjamins.

Chui, Kawai. 2018. Directionality of change: Grammatical variation and Do-constructions in Taiwan Mandarin. *Concentric: Studies in Linguistics*, 44(1):65–88.

Conrad, Susan and Douglas Biber. 2009. *Real Grammar: A Corpus-Based Approach to English*. Pearson Longman, London, UK.

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach.* London: Longman.

Croft, William. 2014. Comparing categories and constructions crosslinguistically (again): The diversity of ditransitives. *Linguistic Typology*, 18(3):533–551.

Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.

Cuervo, Maria Cristina. 2010. Two types of (apparently) ditransitive light verb constructions. In Karlos Arregi, Zsuzsanna Fagyal, Silvina Montrul and Annie Tremblay, editors, *Romance Linguistics 2008: Interactions in Romance. Selected papers from the 38th Linguistic Symposium on Romance Languages (LSRL)*, Urbana-Champaign, April 2008. John Benjamins, Amsterdam, pages 139–156.

Evteeva, Mariya Yur'evna. 2017. On semantic opposition “VERBS vs TO MAKE + VERBAL NOUN”. *Philology. Theory & Practice*, 10(76), Part 1:104–108.

Feltgen, Quentin, Benjamin Fagard, and Jean-Pierre Nadal. 2017. Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science*, 4(11), 170830.

Ghanbarnejad, Fakhteh, Martin Gerlach, José M. Miotto and Eduardo G. Altmann. 2014. Extracting information from S-curves of language change. *J. R. Soc. Interface*, 11: 20141044.

Goddard, Cliff and Anna Wierzbicka. 2002. Semantic Primes and Universal Grammar. In Cliff Goddard and Anna Wierzbicka, editors, *Meaning and Universal Grammar: Theory and Empirical Findings, Volume I*, John Benjamins, Amsterdam, pages 41–85.

Goddard, Cliff. 2002. The search for the shared semantic core of all languages. In C. Goddard, & A.

Wierzbicka, editors, *Meaning and Universal Grammar - Theory and Empirical Findings, Volume I,* John Benjamins, Amsterdam, pages 5–40.

Goddard, Cliff and Anna Wierzbicka. 2007. Semantic primes and cultural scripts in language learning and intercultural communication. In Farzad Sharifian and Gary B. Palmer, editors, *Applied Cultural Linguistics: Implications for Second Language Learning and Intercultural Communication*, John Benjamins, Amsterdam, pages 105–124.

Goddard, Cliff and Anna Wierzbicka. 2014. Semantic fieldwork and lexical universals. *Studies in Language*, 38(1):80–127.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL.

Good, Jeff. 2008. *Linguistic Universals and Language Change*. Oxford: Oxford University Press.

Hernández, Roberto Mayoral. 2007. A variation study of verb types and subject position: Verbs of light and sound emission. In José Camacho, Nydia Flores-Ferrán, Liliana Sánchez, Viviane Déprez and María José Cabrera, editors, *Romance Linguistics 2006: Selected papers from the 36th Linguistic Symposium on Romance Languages (LSRL)*, New Brunswick, March-April 2006. John Benjamins, Amsterdam, pages 213–226.

Hernández, Roberto Mayoral. 2008. Mapping semantic spaces: A constructionist account of the "light verb" xordæn 'eat' in Persian. In Martine Vanhove, editor, *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*. John Benjamins, Amsterdam, pages 139–161.

Hruschka, Daniel J., Morten H. Christiansen, Richard A. Blythe, William Croft, Paul Heggarty, Salikoko S. Mufwene, Janet B. Pierrehumbert and Shana Poplack. 2009. Building social cognitive models of language change. *Trends in Cognitive Science*, 13(11):464–469.

Huang, Chu-Ren and Jingxia Lin. 2012. The ordering of Mandarin Chinese light verbs. In Donghong Ji and Guozheng Xiao, editors, *Chinese Lexical Semantics: The 13th Chinese Lexical Semantics Workshop*. Springer, Heidelberg, pages 728–735.

Huang, Chu-Ren, Jingxia Lin, Menghan Jiang and Hongzhi Xu. 2014. Corpus-based study and identification of Mandarin Chinese light verb variations. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 1–10, Association for Computational Linguistics and Dublin City University. Dublin, Ireland.

Hull, David L. 1988. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, Chicago, IL.

Hull, David L. 2001. *Science and Selection: Essays on Biological Evolution and the Philosophy of Science*. Cambridge University Press, Cambridge, UK.

Kettnerová, Václava and Markéta Lopatková. 2020. Reciprocity in Czech light verb constructions: The dependency perspective. *Jazykovedný Časopis*, 71(1):41–68.

Kettnerová, Václava. 2021. Optional valency complementations in Czech light verb constructions. *Linguistica Pragensia*, 31(1):7–27.

Kovalevskaite, Jolanta, Erika Rimkutė and Laura Vilkaitė-Lozdienė. 2020. Light verb constructions in Lithuanian: Identification and classification. *Studies about Languages*, 36:5–16.

Maiko, Tatsiana. 2019. Конструкции с опорным глаголом в речи изучающих русский язык как иностранный (Support (light) verb constructions in the speech of learners of Russian as a foreign language). In Iliyana Krapova, Svetlana Nistratova and Luisa Ruvoletto, editors, *Studi di linguistica slava. Nuove prospettive e metodologie di ricercar*, pages 285–301.

Maiko, Tatsiana. 2020. What can you give in Italian that you can’t give in Russian? A contrastive study of constructions with the light verbs *dare* in Italian and *davat'/dat'* in Russian. In Joanna Szerszunowicz and Martyna Awier, editors, *Reproducible Multiword Expressions from a Theoretical and Empirical Perspective*, pages 33–54.

Martínez Linares, María Antonia. 2013. Light verb constructions in Latin American newspapers: Creative variants and coinages. *Spanish in Context*, 10(1):114–135.

Nolan, Brian. 2015. Determining light verb constructions in contemporary British and Irish English. *International Journal of Corpus Linguistics*, 20(3):326–354.

Ong, Christina Sook Beng. 2021. Nativised structural patterns of make light verb construction in Malaysian English. *Concentric: Studies in Linguistics*, 47(1):93–112.

Radimský, Jan. 2010. *Verbo-nominální predikát s kategoriálním slovesem (Verb-noun Predicates with a Light Verb)*. Jihočeská univerzita, České Budějovice.

Ronan, Patricia and Gerold Schneider. 2017. Spanish infinitives borrowed into Zapotec light verb constructions. In Karen Dakin, Claudia Parodi and Natalie Operstein, editors, *Language Contact and Change in Mesoamerica and Beyond*. John Benjamins, Amsterdam, pages 56–80.

Ronan, Patricia. 2014. Complex predicates and light verb constructions in Modern Irish. *Revista Española de Lingüística Aplicada/Spanish, Journal of Applied Linguistics*, 27(1):140–167.

Sanromán Vilas, Begoña. 2019. Light verb constructions in spoken L2 English: An exploratory cross-sectional study. In Vaclav Brezina, Dana Gablasova and Tony McEnery, editors, *Corpus-based Approaches to Spoken L2 Production: Evidence from the Trinity Lancaster Corpus*. John Benjamins, Amsterdam, pages 181–206.

Sundquist, John D. 2020. Give as a light verb. *Functions of Language*, 27(3):280–306.

Tadao, Miyamoto. 2000. *The Light Verb Construction in Japanese: The Role of the Verbal Noun*. John Benjamins, Amsterdam.

Wittenberg, Eva and Maria Mercedes Piñango. 2011. Processing light verb constructions. *The Mental Lexicon*, 6(3):393–413.

Yim, Changguk. 2020. Productivity, richness, and diversity of light verb constructions in the history of American English. *Journal of Historical Linguistics*, 10(3):349–388.

Appendix. Abbreviations of linguistic terms.

CL	classifier
COP	copula
DE	morpheme *de*
GEN	genitive
IPFV	imperfective aspect
PFV	perfective aspect
PL	plural
PRF	perfective morpheme
PRS	present tense
PST	past tense
SG	singular
1	first person
2	second person
3	third person

Exploring Metaphorical Polysemy with Multiple Correspondence Analysis: A Corpus-based Study on the Predicative *hēi* ‘black’ in Chinese

Jinmeng Dou
Department of Linguistics and Translation,
City University of Hong Kong,
Hong Kong, People’s Republic of China
jmdou2-c@my.cityu.edu.hk

Meichun Liu
Department of Linguistics and Translation,
City University of Hong Kong,
Hong Kong, People’s Republic of China
meichliu@cityu.edu.hk

Abstract

This paper provides a corpus-based, statistical analysis to explore the semantic (dis)similarities of four metaphorical meanings of the Chinese color term *hēi* ‘black’ regarding its predicative usages. With the Behavioral Profiles approach, 379 instances were manually annotated with 35 contextual features proposed from three categories, including lexical-collocational, morphosyntactic and semantic, and discourse information, to capture the contextual variations of *hēi*. Based on the annotated data, the Multiple Correspondence analysis (MCA) technique is then used to visualize the semantic distribution of the four meanings of *hēi* and their strength of associations with the distinctive features. It is found that the semantic (dis)similarities of the four meanings of *hēi* are well profiled by the MCA results, which demonstrates the effectiveness of the MCA method in studying metaphorical polysemy.

1 Introduction

As a common source of metaphor, Color Terms (CTs) possess a diverse range of metaphorical meanings in various languages beyond their literal meanings pertaining to the natural color, which can be viewed as metaphorical polysemy (Apresjan, 1974; Jurafsky, 1996). Without exception, Chinese CTs also show metaphorical polysemy with varied usage patterns. For example, one of the earliest-acquired Chinese CTs, *hēi* 黑 ‘black’ (Berlin & Kays, 1969; Wu, 2011), is continuously extended to different metaphorical meanings with flexible contextual variations. Except for the attributive uses (e.g., *hēi-shèhuì* 黑社会 ‘black-society: underground’), *hēi* frequently occurs as a predicate when referring to metaphorical meanings. Example (1) below provides initial illustrations for the metaphorical usages:

(1) a. 她无意中黑了青岛一把。
tā wúyìzhōng hēi le qīngdǎo yī bǎ
she unconsciously black LE Qingdao one-CL
‘She accidentally blackened Qingdao for once.’

b. 穆司爵脸黑了。
Mù-sījué liǎn hēi le
NAME face black LE
‘Mu Sijue’s face blackened.’

As shown in (1), the predicate *hēi* can either serve as a transitive verb to denote ‘an action that damages or destroys someone’s reputation’ as in (1a), or an intransitive verb ‘to get angry’ as in (1b). It is noteworthy that the contexts of *hēi* referring to different metaphorical meanings may vary as they correlate with the possible semantic (dis)similarities regarding Distributional Semantics (Harris, 1954; Firth, 1951). However, most previous studies discussed the possible meanings of CTs from a purely analytical ground and regarded the different senses as discrete primes, ignoring their potential semantic relations. Given that, this study aims to explore the semantic (dis)similarities of the varied metaphorical meanings of *hēi* as a predicate with the corpus-based Behavioral Profiles (BP) approach.

The paper is organized as follows. Section 2 offers a literature review of relevant previous studies. Section 3 introduces the research methodology. Section 4 applies the MCA method to identify the distinctive contextual variations of each metaphorical meaning denoted by the predicative *hēi* based on the annotated distinctive contextual features. Section 5 concludes this study.

2 Literature Review

While polysemy is a universal phenomenon in languages (Copestake & Briscoe, 1995; Jackendoff, 2002; Murphy, 2002; Pustejovsky, 1995), metaphorical polysemy is regarded as a special type of polysemy (Apresjan, 1974). Metaphorical polysemy of CTs has been demonstrated in many languages, such as English (Allan, 2009), European languages (Hill, 2008), and Persian (Amouzadeha et al., 2011; Aliakbari & Khosravian, 2013). Besides, several translation-related studies also suggested that some metaphorical meanings of CTs are language-specific or culturally exclusive (Wierzbicka, 1990; Ghafel & Mirzaie, 2014; Chatti, 2016; Hastürkoğlu, 2018; Al-Jarf, 2019).

With regard to Chinese CTs, some studies have discussed their metaphorical extensions and the underlying cognitive mechanisms, such as Wu (1986), Zhang (1988), Xing (2008), Li and Bai (2013), Lai and Chung (2018), etc. Nevertheless, most of them are qualitative analyses based on intuitive judgments with elicited examples without considering their contextual variations or the relations between different senses. As one of the Chinese basic CTs, *hēi* continues to be extended over time to derive various metaphorical meanings, showing multiple form-meaning mapping relations in its metaphorical uses. Specifically, the intransitive verbal uses of *hēi* can either refer to 'being malevolent' (e.g., *xīnhēile* 心黑了 'heart blackened') or 'being angry' (e.g., *liǎnhēile* 脸黑了 'face blackened') depending on the collocated subjects. Besides, the frequent occurrence of the predicative *hēi* with various metaphorical extensions can be barely found in other Chinese CTs, indicating distinct cognitive mechanisms of this term. Hence, this study selects the different metaphorical meanings of the predicative *hēi* as the research object to explore the form-meaning correlations with the BP approach.

Behavioral Profiles (Divjak & Gries, 2009; Gries & Divjak, 2009), as a corpus-based approach, combines the analysis of manually annotated usage features, also named ID tags, (Gries, 2006) with multifactorial statistical tools. In this study, Multiple Correspondence analysis (MCA) (Benzécri, 1992; Greenacre & Blasius, 2006) is adopted to conduct the statistical feature analysis. Pertaining to correspondence analysis, MCA can simultaneously summarize and visualize the correlations between multiple linguistic features that structure the behaviors of the datasets in relation to syntax, semantics, or pragmatics by showing their relative proximity in a biplot. Previous studies have demonstrated the effectiveness of MCA in analyzing lexical semantic issues, e.g., Glynn (2014a) on several mental and communicative predicates in English, Glynn (2016) on the polysemy of the verb *annoy*, Krawczak and Glynn (2015) on three English constructions, as well as Tantucci and Wang (2020) on the aspect marker *guo* and the sentence-final particle *ba* in Chinese.

Furthermore, some significant effects of metaphor have been detected in depicting the semantic (dis)similarities of polysemous lexemes in previous studies. Gries (2006) argued that the metaphorical mappings identified from the polysemous extensions can strongly increase the predictive power of sense recognition and the descriptive power of the clustering algorithm. Besides, other BP studies also indicated the predictive power of metaphor in profiling the semantic relations of polysemy, e.g., Glynn (2014b), Jansegers et al. (2015), Jansegers and Gries (2017), and Ioannou (2020).

3 Data Preparation

3.1 Data Collection

Based on proposals of previous studies and the Contemporary Chinese Dictionary (7th ed.), four metaphorical meanings commonly associated with the predicative *hēi* were selected for the BP analysis, as detailed in Table 1. Consequently, 379 instances pertaining to the four meanings of the predicative *hēi* (94 for "Slander/Entrap", 90 for "Evil/Malevolent", 95 for "Angry/Sullen", and 100 for "Network Attack") were randomly collected from two corpora in Sketch Engine - the Corpus of Chinese Simplified Web 2017 Sample and the Chinese Gigaword 2 Corpus (Mainland, simplified), which guarantee a relatively balanced coverage as the former is composed of internet texts and the latter newswires.

Senses & Examples
1. Slander/Entrap 赵丽颖还一直被黑。 *Zhào-lìyǐng hái yīzhí bèi hēi* NAME still always PASSIVE black ‘Zhao Liying is still always blackened.’
2. Evil/Malevolent 这里的人心太黑。 *zhèlǐ de rén xīn tài hēi* here DE people heart too black ‘The hearts of the people here are too black.’
3. Angry/Sullen 齐王的脸更黑了。 *Qí-wáng de liǎn gèng hēi le* King-Qi DE face more black LE ‘King Qi's face is even darker.’
4. Network Attack 这家网站已经被黑。 *zhè-jiā wǎngzhàn yǐjīng bèi hēi* this-CL website already PASSIVE black ‘This website has been hacked.’

Table 1: Four metaphorical senses pertaining to the predicative *hēi*

3.2 Data Annotation

In line with previous studies (Gries, 2006; Gries & Divjak, 2009; Liesenfeld et al., 2020), a total of 35 contextual features containing 99 variable levels fall into three categories: the lexical-collocational patterns (24), morphosyntactic and semantic behaviors (4), and discourse information (7), as shown in Table 3. For the lexical-collocational patterns, the selected features pertain to the collocations of *hēi* with other lexical categories, such as Degree Markers. The morphosyntactic and semantic features mainly concern the syntactic categories of *hēi* and the semantic types of its surrounding arguments in the same constituent, e.g., the POS of *hēi*, the semantic type of subjects collocated with *hēi*. The discourse features refer to the contextual information beyond phrasal structures, including the functional types of the clause containing *hēi*, the mood, etc.

Instance	Tag 1	Tag 2	…	Tag *j*
Instance 1	C_{11}	C_{12}		C_{1j}
Instance 2	C_{21}	C_{22}		C_{2j}
…	…	…		…
Instance *i*	C_{i1}	C_{i2}		C_{ij}

Table 2. Data format for the annotation

The 35 contextual features were then manually annotated on 379 instances, which produced a data frame of 37,521 data points. The annotation was performed (Table 2) by two native speakers who are linguistically well-trained. Moreover, Cohen’s Kappa coefficient was adopted to measure the inter-rater reliability. Based on the annotation of the two annotators, a high degree of Cohen’s Kappa (0.9066, greater than the threshold value 0.8) was obtained, indicating an almost perfect strength of agreement between the two annotators (Landis & Koch, 1977; Altman, 1991; McHugh, 2012). Then, the annotations of the remaining inconsistent cases were determined by the two annotators after a discussion.

Based on the annotated data, the MCA was conducted by means of the *FactoMiner* package in R language.

Feature Type	Feature	Feature Levels
1. Lexical-Collocational Information		
modifier	Negation Marker	2: yes/no
modifier	Degree Marker	2: yes/no
modifier	*bèi* 被	2: yes/no
modifier	*yě* 也	2: yes/no
modifier	*qǐ* 起	2: yes/no
modifier	*le* 了	2: yes/no
modifier	*zhe* 着	2: yes/no
modifier	*dōu* 都	2: yes/no
modifier	*hái* 还	2: yes/no

modifier	*jiù* 就	2: yes/no
modifier	*guò* 过	2: yes/no
modifier	*ràng/lìng/shǐ* 让/令/使	2: yes/no
modifier	Past Time Marker	2: yes/no
modifier	Future Time Marker	2: yes/no
modifier	Comparison Marker	2: yes/no
modifier	Frequency/Duration Marker	2: yes/no
modifier	Capability/Intention Marker	2: yes/no
modifier	*yuè/yù* 越/愈	2: yes/no
modifier	Doubt Marker	2: yes/no
hēi	The freq. of *hēi* exceeds one time	3:no/yes diff M/yes same M
color terms	Cooccurred with other color terms	2: yes/no
noun phrase	Color object	2: yes/no
notional word	parallel with *hēi* (e.g., 黑恶 black-evil)	3: no/yes diff M/yes same M
compound word	Phase marker	2: yes/no
2. Morphosyntactic & Semantic Information		
POS	part of speech of *hēi*	2: verb/adjective
noun phrase	Semantic type: dep-relation: sub	6: see notes
noun phrase	Semantic type: dep-relation: dobj	6: see notes
verb phrase	Semantic type: N collocated with *hēi*	6: see notes
3. Discoursal Information		
clause	Clause Type	2: main/dependent
clause	Types of dependent clause	3: rel. clause/adv. clause/null
sentence	The omission of co-arguments with *hēi*	2: yes/no
sentence	Pronouns	2: yes/no
sentence	Explication of *hēi*	2: yes/no
pragmatic	Whether *hēi* is ambiguous	2: yes/no
pragmatic	Mood	3: see notes

Table 3: An overview of the proposed 35 contextual features. Notes: 1) The semantic types consist of abstract entity, body part, animate, inanimate object, organization, and null; 2) The sentence mood consists of declarative, interrogative, and imperative.

4 MCA results

4.1 Semantic Distribution

This section discusses the semantic distribution of the four meanings via visualizing the 379 annotated instances with the MCA map. Basically, MCA describes the obtained variations between the categorical variables through several dimensions and uses the so-called 'Inertia' to represent the proportion of variation retained by each dimension. The higher the inertia one obtains, the better the dimension is. Table 4 provides an overview of the inertia of our MCA result on the annotated dataset. It is shown that the variance of the dataset is decomposed into 47 dimensions. Each of them occupies comparatively a value of inertia and thus explains the percentage of the total variation in the data. The reported low score is to be expected regarding the high complexity of our dataset.

Dim.	Eigenvalue	Inertia	Cum. Inertia
1	0.095	6.905	6.905
2	0.072	5.231	12.135
3	0.066	4.847	16.982
4	0.058	4.199	21.181
……			
47	0.004	0.267	100.000

Table 4: Inertia of MCA results for the four meanings of *hēi*

In line with previous studies, the inter-individual variability of the 379 instances is displayed based on the first two dimensions (12.14%) as in Figure 1. The data points (instances) in this figure are colored according to their corresponding meanings of *hēi* and labeled based on their sequential number (1-94 for “Slander/Entrap”, 95-184 for “Evil/Malevolent”, 185-279 for “Angry/Sullen”, 280-379 for “Network Attack”). The positions of the four metaphorical meanings are predicated with 95% confidence ellipses by regarding them as supplementary variables.

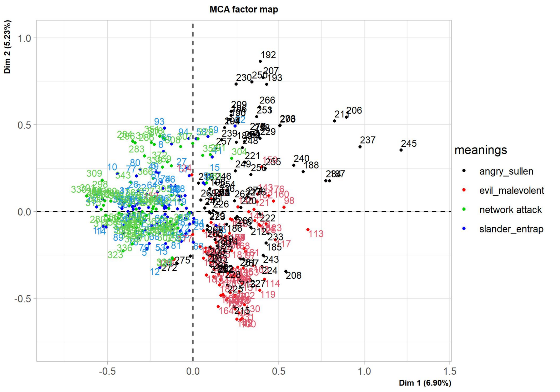


Figure 1: Semantic Distribution of the four meanings of *hēi*

In Figure 1, the two left quadrants are dominated by the blue (“Slander/Entrap”) and green (“Network Attack”) data points, which are overlapped without a distinct boundary. Related to that, their respective confidence ellipses are right next to each other, indicating that these two meanings are strongly associated with each other in semantic properties. On the other hand, the black data points for “Angry/Sullen” are mainly distributed in the top-right quadrant, and the red points for “Evil/Malevolent” in the bottom-right quadrant. Their data points are partially overlapped, which form a fuzzy boundary. It is shown that the usage patterns of these two meanings are semantically distinct, but still share some common properties to a certain extent. Lastly, distinct semantic dissimilarities can be detected between the two right-quadrant meanings and the two left-quadrant meanings, as their data points are divided into two groups by the *y*-axis in Figure 1.

In sum, the semantic (dis)similarities of the four metaphorical meanings of *hēi* are well illustrated in Figure 1, based on their semantic distributions. In the following sections, we focus on identifying the distinctive usage patterns of the varied meanings to provide explanations for their semantic (dis)similarities.

4.2 Distinctive Contextual Features

To analyze the usage patterns of the four meanings, the top 10 distinctive contextual features that set them apart were identified first, since MCA is not suitable for representing too many features simultaneously due to the difficulty of visualization. Precisely, Cramér's V (Cramér, 1946) was calculated to measure the distinctiveness of the contextual features. The Fisher exact test was used to see whether the Cramér's V is statistically significant. Table 5 lists the top 10 distinctive contextual features based on their Cramér's V and *p*-value.

Feature	Cramér's V	*p*-value
POS	0.862	< 0.01
bei 被	0.594	< 0.01
Omission of co-arg	0.593	< 0.01
Semantic type of dobj	0.522	< 0.01
Degree marker	0.478	< 0.01
Semantic type of sub	0.454	< 0.01
zhe 着	0.346	< 0.01
Semantic type of co-N	0.339	< 0.01
Explication	0.279	< 0.01
Color object	0.264	< 0.01

Table 5: Top 10 distinctive contextual features in setting the four meanings apart

In terms of feature types, four of them pertain to lexical-collocational patterns, including passive marker *bèi*, degree markers, aspect marker *zhe*, and color object; four pertain to morphosyntactic and semantic behaviors, including POS of *hēi*, semantic types of co-arguments and other nouns collocated with *hēi*; and two of them refer to discourse information, which are the omission of co-arguments and the explication of *hēi*. It is clear that all three feature types play a crucial role in distinguishing the four meanings.

4.3 Distinctive usage patterns of the four meanings

The strength of associations between the four meanings and the feature categories was visualized with the MCA method based on the distinctive features. The cumulative inertia of the first two dimensions (Table 6) was used to plot the MCA factor map, as shown in Figure 2. Only the data points with a cos2 value (quality of representation) exceeding 0.05 were labeled in Figure 2 to avoid misguiding by the underrepresented points.

Dim.	Eigenvalue	Inertia	Cum. Inertia
1	0.295	15.580	15.580
2	0.205	10.792	26.373
3	0.175	9.219	35.592
4	0.132	6.949	42.541
5	0.121	6.392	48.932
……			
19	0.015	0.789	100.000

Table 6: Inertia of MCA results for the top 10 distinctive features

In Figure 2, the data points of contextual features are colored in red, and the positions of the four meanings of *hēi* are predicated in green as supplementary variables. As mentioned above, the strength of associations between the contextual features and the four meanings is visualized as their proximity to each other. Based on Figure 2, what follows is a detailed analysis on the usage patterns of these four metaphorical meanings.

In Figure 2, the two meanings "Slander/Entrap" and "Network Attack" are observed on the left quadrants near the x-axis, while the other two meanings, "Angry/Sullen" and "Evil/Malevolent", are located in the right quadrants with relatively farther distance. As indicated by features "bei_yes", "STO_Org" and "STO_Abs", *hēi* tends to be a transitive verb that takes an object pertaining to an organization or abstract entity in a passive construction, when referring to "Network Attack". Moreover, its subject-agent is more likely to be omitted based on features "STS_no" and "OmCA_yes". Example (2) below illustrates this pattern.

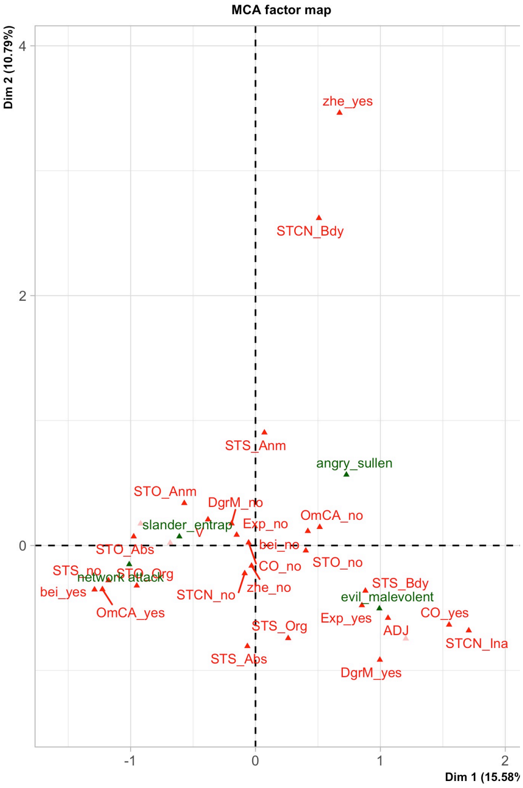


Figure 2: Associations of the four meanings with the distinctive contextual features

(2) 全球有 8 万家公司被黑。
quánqiú yǒu 8-wàn-jiā gōngsī bèi hēi
world have 80,000-CL company PASSIVE black
'There are 80,000 companies worldwide were hacked.'

For meaning "Slander/Entrap", *hēi* may also behave as a transitive verb with an animate subject and an animate or abstract object, as indicated by features "V", "STS_Anm", "STO_Anm" and "STO_Abs". Example (3) illustrates this usage.

(3) 你还会黑她吗？
nǐ hái huì hēi tā ma
you still can black she MA
'Will you still slander her?'

On the other hand, the data point of "Evil/Malevolent" distributes in the bottom-right quadrant with a group of contextual features pertaining to the uses of adjectival predicate. Precisely, it is shown that *hēi* referring to this meaning behaves predominately as a stative predicate collocating with subjects referring to body part, based on features "ADJ" and "STS_Bdy". Besides, a degree marker or an inanimate color object is frequently observed in such uses to describe the gradation of the black color, as indicated by the features “DgrM_yes”, “CO_yes” and “STCN_Ina”. Example (4) illustrates this usage. Regarding discourse information, the feature “Exp_yes” indicates that the reason why the referred subject is black may also be expressed in the context.

(4) 他的心[很黑/黑得和碳一样]。
tāde xīn [*hěn hēi*/*hēi dé hé tàn yīyàng*]
his heart [very black/black DE and carbon same]
‘His heart is very black (black as carbon).’

For the meaning “Angry/Sullen”, it is found that *hēi* is more likely to be an intransitive verbal predicate of which subjects refer to human or body part. In other words, the features “V” and “STS_Anm” are attracted by the meaning “Angry/Sullen”, which explains their relative distance to the two left-quadrant meanings. Example (5) explains this usage.

(5) 他的脸瞬间黑了。
tāde liǎn shùnjiān hēi le
his face instantly black LE
‘His face blackened immediately.’

In addition, features “zhe_yes” and “STCN_Bdy” show a strong attraction to “Angry/Sullen”, as *hēi* in this meaning frequently collocates with durative aspect marker *zhe* and nouns pertaining to body part. This usage is illustrated by (6).

(6) 他黑着脸说话.
tā hēi zhe liǎn shuōhuà
he black ZHE face speak
‘He speaks with a black face.’

In sum, it is found that when referring to “Network attack” and “Slander/Entrap”, *hēi* prototypically serves as a causative-transitive verb to describe the hostile social or interpersonal interactions; when referring to “Evil/Malevolent”, *hēi* tends to be a stative predicate to denote the evaluation on the immoral quality of a human or social entity; when referring to “Angry/Sullen”, *hēi* may serve as an intransitive verbal predicate to depict an unpleasant emotional status.

5 Conclusion and Implication

In this paper, a corpus-based BP analysis was conducted to explore the semantic (dis)similarities in relation to metaphorical polysemy of the predicative uses of the Chinese CT *hēi*. It is found that as a transitive predicate, *hēi* tends to describe an action of social attack as blackening someone or something, with the meanings “Slander/Entrap” or “Network Attack”. From the perspective of Conceptual Metaphor, the findings suggest that a conceptual mapping may be at work, which explains the association of the two meanings: (i) “SOCIAL ATTACK IS BLACKENING”. As an intransitive, stative predicate, *hēi* prototypically depicts a negative evaluation on a social entity, with the meaning “Evil/Malevolent”. A conceptual mapping can be postulated for this meaning as: (ii) “EVALUAITION OF IMMORAL QUALITY IS PERCEPTION OF THE BLACK COLOR”. Regarding the meaning “Angry/Sullen”, *hēi* mainly describes the process of turning into an unpleasant mental state, with the conceptual metaphor (iii) “TURNING ANGRY IS TURNING BLACK”. Given the proposed metaphors, the semantic (dis)similarities of the four meanings can be well accounted for based on their entailed conceptual mappings, which can be specified as: “change of social or emotional status corresponds to change of color” for metaphors (i) & (iii), “non-visual evaluation of quality or mental state corresponds to color perception” for metaphors (ii) & (iii).

Ultimately, this study demonstrates the effectiveness of the BP approach with MCA method in exploring metaphorical polysemy, with empirical evidence that can be visually represented.

References

Andreas Liesenfeld, Meichun Liu, and Churen Huang. 2020. Profiling the Chinese causative construction with *rang* (讓), *shi* (使) and *ling* (令) using frame semantic features. *Corpus Linguistics and Linguistic Theory*, 18(2):263-306.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1):15-67.

Anna Wierzbicka. 1990. The meaning of color terms: Semantic, culture and cognition. *Cognitive Linguistics*, 1(1):99-150.

Banafsheh Ghafel and Akbar Mirzaie. 2014. Colors in everyday metaphoric language of Persian speakers. *Procedia - Social and Behavioral Sciences*, 136:133-143.

Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. University of California Press.

Dagmar Divjak and Stefan Th. Gries. 2009. Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In Katarzyna Dziwirek and Barbara Lewandowska-Tomaszczyk (Eds.), *Studies in cognitive corpus linguistics* (pp. 273-296). Peter Lang.

Daniel Jurafsky. 1996. Universal tendencies in the semantics of the diminutive. *Language*, 72(3):533-578.

Douglas G. Altman. 1991. *Practical statistics for medical research*. Chapman and Hall Press.

Dylan Glynn. 2014a. Correspondence analysis: Exploring data and identifying patterns. In Dylan Glynn and Justyna A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 443-485). John Benjamins.

Dylan Glynn. 2014b. The many uses of *run*: Corpus methods and Socio-Cognitive Semantics. In Dylan Glynn and Justyna A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 117-144). John Benjamins.

Dylan Glynn. 2016. Quantifying polysemy: Corpus methodology for prototype theory. *Folia Linguistica*, 50(2):413-447.

Georgios Ioannou. 2020. Image schemas as prototypes in the diachronic evolution of *kámnō* and *eutheiázō* in Greek: A behavioural-profile analysis. *Lingua*, 245:1-32.

Gökçen Hastürkoğlu. 2018. Incorporation of conceptual metaphor theory in translation pedagogy: A case study on translating simile-based idioms. *Australian Journal of Linguistics*, 38(4):467-483.

Gregory Murphy. 2002. *The big book of concepts*. MIT Press.

Harald Cramér. 1946. *Mathematical methods of statistics*. Princeton University Press.

Hueiling Lai and Siawfong Chung. 2018. Color polysemy: black and white in Taiwanese language. *Taiwan Journal of Linguistics*, 16(1):95-130.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical Data. *Biometrics*, 33(1):159-174.

James Pustejovsky. 1995. *The generative lexicon*. MIT Press.

Janet Zhiqun Xing. 2008. Semantics and pragmatics of color terms in Chinese. In Janet Zhiqun Xing (Ed.), *Studies of chinese linguistics: Functional approaches* (pp. 87-102). Hong Kong University Press.

Jean-Paul Benzécri. 1992. *Correspondence analysis handbook*. Marcel Dekker.

Jianshe Wu. 2011. The evolution of basic color terms in Chinese. *Journal of Chinese Linguistics*, 39(1):76-122.

John R. Firth. 1951. *Modes of meaning*. Oxford University Press.

JU. D. Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5-32.

Karolina Krawczak and Dylan Glynn. 2015. Operationalizing mirativity: A usage-based quantitative study of constructional construal in English. *Review of Cognitive Linguistics*, 13(2):353-382.

Keith Allan. 2009. The connotations of English color terms: color-based X-phemisms. *Journal of Pragmatics*, 41(3):626-637.

Marlies Jansegers and Stefan Th. Gries. 2017. Towards a dynamic behavioral profile: A diachronic study of polysemous *sentir* in Spanish. *Corpus Linguistics and Linguistic Theory*, 16(1):145-187.

Marlies Jansegers, Clara Vanderschueren, and Renata Enghels. 2015. The polysemy of the Spanish verb sentir: A behavioral profile analysis. *Cognitive Linguistics*, 26(3):381-421.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276-82.

Michael Greenacre and Jorg Blasius. 2006. *Multiple Correspondence analysis and related methods*. Chapman & Hal/CRCl.

Mohammad Aliakbari and Fereshteh Khosravian. 2013. A corpus analysis of color-term conceptual metaphors in Persian proverbs. *Procedia - Social and Behavioral Sciences*, 70:11-17.

Mohammad Amouzadeha, Manouchehr Tavangar, and Mohammad A. Sorahia. 2011. A cognitive study of color terms in Persian and English. *Procedia - Social and Behavioral Sciences*, 32:238-245.

Peter Hill. 2008. The metaphorical use of colour terms in the Slavonic languages. In David N. Wells (Ed.), *Themes and variations in Slavic languages and cultures* (pp. 62-83). Australian Contributions to the XIV International Congress of Slavists Perth: ANZSA.

Ray Jackendoff. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Reima Al-Jarf. 2019. Translation students' difficulties with English and Arabic color-based metaphorical expressions. *Fachsprache*, 41(S1):101-118.

Sami Chatti. 2016. Translating colour metaphors: A cognitive perspective. In Mustapha Taibi (Ed.), *New insights into Arabic translation and interpreting* (pp. 161-176). Channel View Publications.

Stefan Th. Gries. 2006. Corpus-based methods and cognitive semantics: The many senses of to run. In Stefan Th. Gries and Anatol Stefanowitsch (Eds.), *Corpora in cognitive linguistics* (pp. 57-99). Muton de Gruyter.

Stefan Th. Gries and Dagmar Divjak. 2009. Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In Vyvyan Evans and Stephanie Pourcel (Eds.), *New directions in Cognitive Linguistics* (pp. 57-75). John Benjamins.

Tieping Wu. 1986. 论颜色词及其模糊性质 [Analysis on color terms and their fuzzy nature]. *Language Teaching and Linguistic Studies*, 2: 88-105.

Vittotio Tantucci and Aiqing Wang. 2020. From co-actions to intersubjectivity throughout Chinese ontogeny: A usage-based analysis of knowledge ascription and expected agreement. *Journal of Pragmatics*, 167:98-115.

Wangxi Zhang. 1988. 色彩词语联想意义初论 [Analysis on the associative meanings of color terms]. *Language Teaching and Linguistic Studies*, 3:112-121.

Zellig S. Harris. 1954. Distributional structure. *Word* 10(2-3):146-162.

Zongcheng Li and Haoren Bai. 2013. Conceptual metaphors of black and white: A corpus-based comparative study between English and Chinese. *Journal of Anhui Agricultural University (Social Science Edition)*, 22(4):92-97.

Sentiment Analysis in Code-Mixed Vietnamese-English Sentence-level Hotel Reviews

Dang Van Thin, Duong Ngoc Hao, Ngan Luu-Thuy Nguyen
University of Information Technology, Ho Chi Minh city, Vietnam
Vietnam National University Ho Chi Minh city, Vietnam
{thindv,haodn,ngannlt}@uit.edu.vn

Abstract

In recent years, there has been an increasing amount of research on code-mixed Sentiment Analysis (SA) tasks due to the evolution of social media platforms in a multilingual society. This paper presents a comprehensive study on the Vietnamese-English code-mixed SA task, including (1) releasing two semi-annotated En-Vi code-mixed datasets; (2) investigating the performance of different machine learning, deep learning, and transformer-based approaches. The experimental results demonstrated that fine-tuning the multilingual sentence-transformer LaBSE (Feng et al., 2022) achieves better performance than the remaining approaches on two of our code-mixed SA datasets. Our work is the first tempt to solve the code-mixed Vietnamese-English SA problem to the best of our knowledge.

1 Introduction

The diversity of discussion platforms, such as forums, e-commerce, and social media, allows users to express their opinions and comments. This growth makes it challenging for individuals and organizations to understand users' aggregated thoughts (Ligthart et al., 2021). Therefore, the task of sentiment analysis has received a lot of attention from the NLP community (Liu and Zhang, 2012).

Code-mixing is the phenomenon of mixing the vocabulary and syntax of two or multiple languages in the sentence (Lal et al., 2019). Due to the rise of multilingual environments, there is an increase in code-mixed written text. Unlike monolingual sentences (e.g., English), code-mixing is very challenging for traditional NLP architectures because of grammatical constructions and spelling mistakes. Therefore, there has been a dramatic increase in code-mixed problems (Pratapa et al., 2018; Rani et al., 2020; Patwa et al., 2020; Chakravarthi et al., 2022).

On the other hand, recent multilingual NLP research has attracted the community's attention on word-level (Ruder et al., 2019) and sentence-level representations (Artetxe and Schwenk, 2019; Feng et al., 2022). Besides, the growth of multilingual Transformer-based language models (Devlin et al., 2019; Conneau et al., 2020) brought benefits to many downstream NLP tasks. Therefore, these representation methods can be effective for code-mixed tasks, especially low-resource languages. This paper presents a study on the code-mixed Vietnamese-English data for the SA task. The reason why we choose the Vietnamese-English language is that most Vietnamese people use English as a second language (Doan et al., 2018). Our main contributions can be summarized as follows:

- We release two code-mixed SA datasets for the Vietnamese and English languages for the hotel domain.
- We investigate the effectiveness of different machine learning and deep learning approach on the code-mixed sentiment analysis task.
- We perform experiments to confirm whether fine-tuning the SOTA pre-trained transformers benefit the code-mixed dataset. The experimental results demonstrated that fine-tuning the LaBSE model (Feng et al., 2022) achieves the best results on our datasets.

2 Related Work

In recent years, there has been an increasing interest in code-mixing or code-switching NLP tasks, including hate speech detection (Rani et al., 2020), Part-of-Speech tagging (Pratapa et al., 2018), Language Identification (Aguilar and Solorio, 2020). Moreover, researchers have shown an increased interest in code-mixing SA, however, most datasets are annotated for high-resource languages such as Spanish-English (Patwa et al., 2020), Hindi-English (Swamy et al., 2022; Hande et al., 2020; Patwa et al., 2020), Malayalam-English (Chakravarthi et al., 2020), Persian-English (Sabri et al., 2021), Dravidian-English (Chakravarthi et al., 2022).

Most recent studies focus on the machine learning approach to address the code-mixed sentiment analysis task. The authors (Hande et al., 2020; Patwa et al., 2020; Chakravarthi et al., 2022) investigated the performance of different machine learning and deep learning methods such as Support Vector Machine (SVM), Naive Bayes (NB), Convolution Neural Network (CNN). Moreover, there are some studies (Pratapa et al., 2018; Singh and Lefever, 2020) that utilized the effectiveness of cross-lingual word embedding approaches to perform code-mixing data. Their experimental results showed that incorporating the multilingual embedding increases the performance of baseline methods. With the development of multilingual language models such as mBERT(Devlin et al., 2019), XLM-R(Conneau et al., 2020), there were a few studies (Younas et al., 2020; Gupta et al., 2021) investigated the effectiveness of fine-tuning pre-trained language models for code-mixed SA datasets.

For Vietnamese language, the most recent works conducted on monolingual SA tasks by fine-tuning pre-trained language models (Nguyen et al., 2020; Truong et al., 2020). Our work is the first tempt to solve the code-mixed Vietnamese and English SA tasks to the best of our knowledge. In this paper, we introduce two Vietnamese-English code-mixed datasets and provide the performance of various benchmark approaches, including classical machine learning (ML) with handcraft features or multilingual sentence representations, deep learning with cross-lingual word embeddings and pre-trained language models.

3 Data Collection

The scarcity of code-mixed sentiment analysis datasets limits current study for low-resource languages. To tackle this research gap, we create two Vietnamese-English code-mixed datasets for the hotel reviews.

The development of our datasets are based on the available annotated SA dataset (Duyen et al., 2014). However, we found that this dataset still has some confusion between polarity classes and contains meaningless sentences. Therefore, we filtered and re-annotated the sentiment polarity label to ensure the dataset's quality. General, the term "code-mixed" refers to placing and mixing of words, phrases, and morphemes of two or more languages in the same sentence[1]. Besides, we notice that people often use English idioms or phrases to express ideas in Vietnamese text. Therefore, we create two datasets in two ways as follows: (1) Translating the noun, verb, and adjective to English in the review (named as WordDataset); (2) Translating the extracted keyword to the corresponding phrase (named as KeyDataset).

To build the WordDataset, we use the Vietnamese POS Tagging[2] tool to extract the Noun, Verb, and Adjective in the dataset to create the bilingual dictionary - where each word in this dictionary is translated to English. However, it is worth noticing that a word can have many different corresponding vocabularies in the target language; therefore, we expand this dictionary by adding its synonyms. For example, the word "khách sạn" can be translated as "hotel", "resort", or "hostel". Notice that this bilingual dictionary is corrected by a human with a language background in the hotel domain. Then, we randomize words in a set of extracted words (noun, verb, adjective) in each sentence to be replaced with a randomly corresponding translation word in the bilingual dictionary. This helps us create a dataset with different code-mixed sentences and is suitable for real applications because users can use different vocabulary in foreign languages.

In addition, instead of replacing essential words in the sentence, we also created another dataset based on translating the main keywords in the review. To extract the keyword, we use KeyBERT[3] based on the monolingual pretrained PhoBERT embedding (Nguyen and Tuan Nguyen, 2020). In order to distinguish to WordDataset, we extract keyphrases with the length ranging from 2 to 4 vocabularies. We also limit the number of top keyphrases in the long review to a value of 5. We build a keyphrase dictionary similar to the above

[1] https://en.wikipedia.org/wiki/Code-mixing

[2] https://github.com/trungtv/pyvi

[3] https://github.com/MaartenGr/KeyBERT

Table 1: Summary statistics for two datasets. Length is the average sentence length, Vocab is the size of the vocabulary.

	N.o sentences per class			Length	Vocab
	Positive	Negative	Neutral		
WordDataset	1981	780	543	14.38	47519
KeyDataset	1981	780	543	14.65	48397

procedure, where each value is translated to the target language and checked manually. Then, we randomly replace the extracted keywords in the review to create the new code-mixed dataset.

To ensure the quality of structural naturalness and lexical diversity in the code-mixed sentence, we conduct a revision process based on multilingual annotators to check and correct the dataset. The pseudocode to create word code-mixed datasets is illustrated in Algorithm 1. The detailed statistics of the two datasets are shown in Table 1.

Algorithm 1 Building Vietnamese-English word code-mixed SA Dataset.

```
Input: Vietnamese SA data: $S = \{s_n\}_{n=1}^{N}$ ; a
       Vi-En dictionaries: $dict = \{k; v\}_{m=1}^{M}$
       where $v = \{v_1, v_2, ..., v_k\}$;
Output: Vi-En code-mixed dataset: $T = \{t_n\}_{n=1}^{N}$
for $n \leftarrow 1...N$ do
    w_rep ← [ ];
    list_rep ← [ ];
    for $m \leftarrow 1...M$ do
        if $k_m$ in $s_n$ then
            w_rep.insert($k_m$);
            list_rep.insert($dict[k_m]$) ;
        end
    end
    L ← $Length(w\_rep)$;
    for $i \leftarrow 1...L$ do
        w ← $random(list\_rep[i])$;
        $s_n \leftarrow replace(w\_rep[i], w)$;
    end
    $t_n \leftarrow s_n$;
end
```

4 Methods

This research aims to investigate the performance of different approaches for Vi-En code-mixed SA. Therefore, we conduct extensive experiments based on various methods, which we explain below.

Classical ML models + handcraft features: As a first baseline, we explore the performance of classical machine learning techniques, including SVM and Multilayer Perceptron. We extract the list of handcraft features (Duyen et al., 2014) and convert them to TF-IDF representation. The handcraft features are used as follows:

- **N-grams**: the bigrams of words is extracted as the features.
- **Important words**: We extract the main words in the review, including noun, verbs and adjectives.
- **POS Information**: All Part-of-Speech of words in the sentences.

Classical ML models + Sentence embedding: We also consider multilingual sentence embedding as the primary representation. In this case, we consider it as feature extraction and train them on ML classifiers. To the best of our knowledge, our study is the first attempt to explore the performance of multilingual embedding for the Vietnamese SA task. To extract the sentence representation, we investigate two newest cross-lingual sentence embeddings as below:

- **LASER**: LASER (Language-Agnostic Sentence Representations) is an encoder to generate pre-trained language representation in 93 languages, including very low-resource languages. It is able to map the sentences with the semantic closeness of different languages in a shared semantic vector space. The detail of LASER's architecture can be seen in the original work (Artetxe and Schwenk, 2019). This model achieved promising results for sentence-level NLP tasks, therefore, it is suitable to extract the code-mixed sentence representation.
- **LaBSE**: A related development is that of Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022). This model is based on a pre-trained BERT-like architecture, and dual encoder models create cross-lingual sentence embedding of 109 languages. The pre-trained model is also released to support downstream NLP tasks. In addition, this model can be used to represent parallel sentences through high-dimensional embeddings.

Deep learning approaches: Cross-lingual word embedding (Ruder et al., 2019) might be one of the interesting ways for the code-mixed problem. For the code-mixed SA tasks, there are some previous studies (Ma et al., 2020; Younas et al., 2020) which applied the deep learning models combined with various pre-trained cross-lingual embeddings such as MUSE (Lample et al., 2018), BPE(Heinzerling and Strube, 2018). As a result, we explore the effectiveness of two deep learning architectures (CNN (Kim, 2014) and LSTM) with different multilingual embeddings (MUSE and BPE) - which produced high performance in several cross-lingual NLP tasks.

- **MUSE**: MUSE is a toolkit that allows us to align the fastText word embeddings (Grave et al., 2018) in a common semantic space. We use pre-trained Vi-En mapping embedding to represent the words in a sentence and is updated during the training.

- **BPEMulti**: This multilingual subword embedding is trained on Wikipedia texts of 275 languages. This subword embedding is trained on a combination of data from multiple languages, and each subword is represented as 300 dimensions.

Transformer-based approaches: The recent development of transformer architectures has brought significant improvements to the NLP field. In the experiments, we consider three multilingual pre-trained language models, including mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), LaBSE (Feng et al., 2022). A short description of the models is given below:

- **mBERT**: Multilingual BERT is trained on 104 highest resource languages in Wikipedia data. Therefore, this model is able to produce cross-lingual representations that allow for fine-tuning the code-mixed sentence.

- **XLM-R**: This model utilized self-supervised training techniques and was trained on filtered Common Crawled data. Therefore, fine-tuning can significantly improve performance on a variety of cross-lingual benchmarks.

- **LaBSE**: LaBSE is introduced as a pre-trained multilingual language model which is trained based on two tasks: masked language modeling and translation language modeling for 109 languages. The experimental results demonstrated the performance of LaBSE in various NLP tasks(Feng et al., 2022). However, the effectiveness of fine-tuning this model has not been investigated for code-mixed tasks.

As the original work (Devlin et al., 2019), we fine-tune the pre-trained language model by putting the final hidden state **h** of [CLS] token as the representation of the code-mixed sentence. Then, a classifier with softmax activation is added to predict the probability of sentiment class **c**:

$$p(c|h) = softmax(Wh) \qquad (1)$$

where **W** is the parameter matrix. The parameters of transformers and matrix **W** is updated during the training process.

5 Experiments

We use the stratified 5-folds cross-validation to report our experiments. The results are measured based on the micro-averaged and macro-averaged F1-score in all our experiments because of the imbalance in classes. Moreover, the weighted F1-score is used on the test set in previous studies (Patwa et al., 2020). We report the experimental results using different evaluation methods to compare the effectiveness of various approaches objectively.

As described in Section 4, we investigate and compare the performance of different models. For the classical ML model, we use the Linear SVM and the two layers MLP. The hyper-parameters of the two models are optimized using a grid search technique. We fix the architecture to 3 convolution layers with different kernel sizes (2,3,4), a dimensionality of 64 units, and ReLU activation for the CNN model. Then, we concatenate the output of global and max pooling features. For the LSTM model, we employ the bidirectional LSTM with 128 units and ReLU activation. To calculate the probability of polarities, we add two feed-forward layers with 300 dimensions, a ReLU activation for the first layer, and the number of classes dimensions and softmax activation for the second layer. We also apply dropout on the word embedding with a rate of 0.5 to prevent the overfitting of the two models. Two models are optimized using the Adam optimizer with a learning rate of 0.001, a batch size of 64, and a number of epochs of 100. For cross-lingual embeddings, we use the pre-trained MulBPE[4] with the size of 1 million vocabularies.

[4] https://bpemb.h-its.org/multi/

For the MUSE embeddings[5], we use the aligned fastText embeddings.

We used Huggingface's Trainer API (Wolf et al., 2020) to implement the transformers architectures, including mBERT[6], XLM-R[7] and LaBSE[8], and the hyperparameters were optimized using the search functionality offered by Trainer API. For the preprocessing component, we applied the same steps as a previous work (Thin et al., 2019).

5.1 Results and Analysis

Table 2 gives an overview of the results for classical ML models with handcraft features and sentence representations, while Table 3 shows the results of deep learning models combined with cross-lingual word embeddings and multilingual transformer-based language models.

As shown in Table 2, it can be noted that training models on LaBSE sentence representation consistently outperform other types of features in both datasets. It is obvious that handcraft features with TF-IDF representation achieved the lowest scores in terms of three F1 scores. One of the reasons for the poor performance of this approach is the sparsity of feature vectors and the diversity of vocabularies in both languages. Also, we observe that the performance of sentence representation is quite competitive with classical ML classifiers. Combining the sentence representation with ML classifier improves results than deep learning with cross-lingual embeddings. The reason might be because of the size of training data when the deep learning models often require more training data to achieve reasonably good performance. Comparing the results of SVM against MLP classifier shows that SVM yields better performance in types of features. Another interesting point is that ML models trained on LaBSE representation perform better than two remainder popular transformers models in terms of Weighted and Macro F1-score in both datasets. Our results demonstrated that LaBSE sentence embedding could produce an adequate representation for the code-mixed sentences. In addition, the results shown that the performance of models in both datasets is different; however, the difference is not significant.

We can also observe that fine-tuning the pre-trained SOTA multilingual sentence transformer achieved the highest scores in all terms of F1-score in the two datasets. As seen in Table 3, the approach based on LaBSE outperformed more consistently than remainder methods in all evaluation metrics. Figure 1 and Figure 2 show the confusion matrix of LaBSE model on two datasets, respectively. We observe that the "neutral" label has the lowest score, while the two other classes achieve better results. The poor performance of the "neutral" class is because there are multiple sentences with opposite polarity; therefore, the label of these samples is annotated as "neutral" in original corpus(Duyen et al., 2014). For example, "khách sạn có vị trí đẹp nhưng nhân viên lễ tân giao tiếp còn kém, đồ ăn sáng thường." (*The hotel has a nice location, but the reception staff are not good at communicating, the breakfast is normal.*). The first phase in the sentence is expressed positively, while the remainder is the negative attitude of the user. That is why the overall sentiment polarity of these sentences is neural.

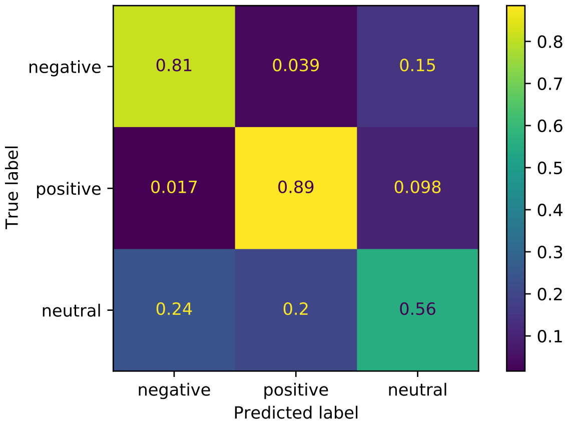


Figure 1: The confusion matrix of the LaBSE model on the WordDataset.

We also conducted an experiment to explore the performance of monolingual language models compared with the multilingual language model in the code-mixed data based on two scenarios on the WordDataset: (1) fine-tuning the latest pre-trained monolingual language models directly on code-mixed data (2) translating the code-mixed sentence to English and Vietnamese, then training translated data using monolingual language models. We choose the base version of PhoBERT (Nguyen and Tuan Nguyen, 2020), and RoBERTa (Liu et al., 2019) with the same above configuration for Vietnamese and English, respectively. We use Google API Translation to translate the code-mixed data

[5] https://github.com/facebookresearch/MUSE
[6] https://huggingface.co/bert-base-multilingual-cased
[7] https://huggingface.co/xlm-roberta-base
[8] https://huggingface.co/sentence-transformers/LaBSE

Table 2: The performance of classical machine learning models on two datasets.

Features	Model	KeyDataset			WordDataset		
		Weighted F1	**Macro F1**	**Micro F1**	**Weighted F1**	**Macro F1**	**Micro F1**
Handcraft + TFIDF	SVM	73.50	62.78	75.64	73.37	62.41	75.27
	MLP	73.60	63.67	74.36	73.39	63.25	73.88
Laser embedding	SVM	74.13	62.39	77.60	74.10	62.20	77.54
	MLP	74.83	65.52	75.27	75.41	66.13	75.48
LaBSE embedding	SVM	77.70	67.54	**79.72**	77.65	67.68	**79.54**
	MLP	**78.49**	**69.81**	78.72	**78.39**	**69.46**	79.21

Table 3: The performance of deep learning and transformers-based models on two datasets.

Approach	Model	KeyDataset			WordDataset		
		Weighted F1	**Macro F1**	**Micro F1**	**Weighted F1**	**Macro F1**	**Micro F1**
Deep learning	CNN + MUSE	75.61	66.47	76.06	74.79	64.86	75.30
	LSTM + MUSE	73.62	63.04	74.43	74.03	63.71	74.15
	CNN + MultiBPE	75.46	65.85	76.24	74.60	64.57	75.24
	LSTM + MultiBPE	73.58	64.11	73.76	73.25	63.24	73.46
Transformers model	mBERT	75.75	63.86	78.66	76.32	65.34	78.42
	XLM-R	76.23	63.59	79.60	76.26	63.98	79.75
	LaBSE	**81.87**	**73.38**	**82.90**	**82.24**	**74.04**	**83.05**

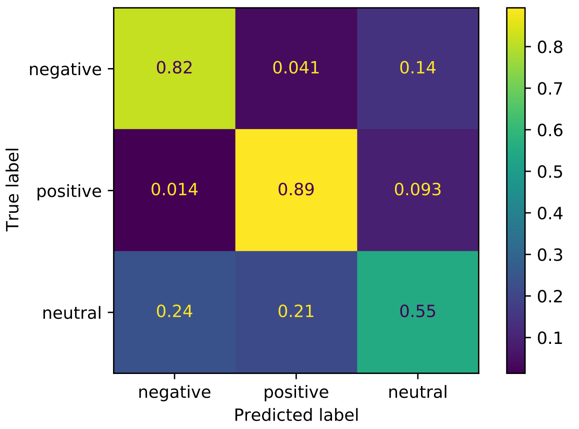


Figure 2: The confusion matrix of the LaBSE model on the KeyDataset.

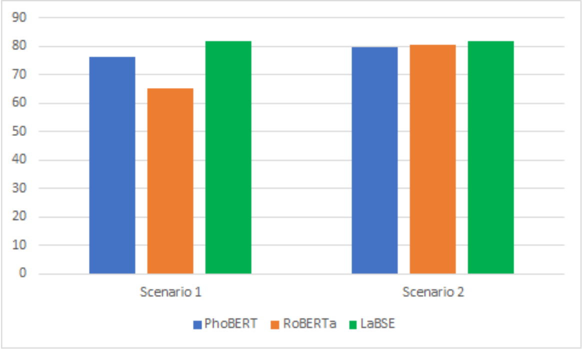


Figure 3: The performance of monolingual models compared with the LaBSE model in two scenarios.

to specific languages (English and Vietnamese) for the second scenario. Figure 3 shows the Weighted F1-score of two scenarios and the LaBSE's performance. It is obvious that fine-tuning directly pre-trained multilingual models gain better results than monolingual models in two scenarios for the code-mixed data. These experiments show that multilingual models are able to achieve competitive results for code-mixed tasks.

6 Conclusion

This paper presents a comprehensive Vietnamese and English code-mixed Sentiment Analysis for the hotel domain. Firstly, we introduced two code-mixed Vi-En datasets created based on the semi-approach. Secondly, we investigated the different methods on two datasets, including the classical ML approach combined with handcraft features or sentence representations, deep learning architectures with cross-lingual word embeddings, and SOTA multilingual language models. It is surprising that fine-tuning the pre-trained LaBSE (Feng et al., 2022) achieved the highest performance. We release two datasets and our code for the research community to facilitate future work on the code-mixed Vi-En SA task.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Gustavo Aguilar and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ngoc Doan, Toan Pham, Min Pham, and Kham Tran. 2018. English as an international language in viet nam: History and development. *Asian Englishes*, 20(2):106–121.

Nguyen Thi Duyen, Ngo Xuan Bach, and Tu Minh Phuong. 2014. An empirical study on sentiment analysis for vietnamese. In *Proceedings of ATC*, pages 309–314.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in Dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 73–79, Kyiv. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. 2019. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 371–377, Florence, Italy. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 54(7):4997–5053.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yili Ma, Liang Zhao, and Jie Hao. 2020. XLP at SemEval-2020 task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 975–980, Barcelona (online). International Committee for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Quoc Thai Nguyen, Thoai Linh Nguyen, Ngoc Hoang Luong, and Quoc Hung Ngo. 2020. Fine-tuning bert for sentiment analysis of vietnamese reviews. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 302–307.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072, Brussels, Belgium. Association for Computational Linguistics.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 42–48, Marseille, France. European Language Resources Association (ELRA).

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Nazanin Sabri, Ali Edalat, and Behnam Bahrak. 2021. Sentiment analysis of persian-english code-mixed texts. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4.

Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 45–51, Marseille, France. European Language Resources Association.

Sowmya Swamy, Jyoti Kundale, and Dipti Jadhav. 2022. Sentiment analysis of multilingual mixed-code, twitter data using machine learning approach. In *Proceedings of ICICC*, pages 683–697.

Dang Van Thin, Duc-Vu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Hoang-Tu Nguyen. 2019. Multi-task learning for aspect and polarity recognition on vietnamese datasets. In *Proceedings of PACLING*, pages 169–180.

Trong-Loc Truong, Hanh-Linh Le, and Thien-Phuc Le-Dang. 2020. Sentiment analysis implementing bert-based pre-trained language model for vietnamese. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 362–367.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Aqsa Younas, Raheela Nasim, Saqib Ali, Guojun Wang, and Fang Qi. 2020. Sentiment analysis of code-mixed roman urdu-english social media text using deep learning approaches. In *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*, pages 66–71.

A Contrastive Corpus-based Analysis of Rhetoric in Asian, American, and European Low-Cost Carriers' Slogans

Ramsey S. Ferrer
Philippine State College of Aeronautics
De La Salle University-Manila
ferrer.ramsey@gmail.com

Abstract

The COVID-19 pandemic has drastically affected various industries, and aviation is the hardest hit (Suau-Sanchez et al., 2020). This predicament has turned all flyers across the globe into flying a low-cost carrier (henceforth, LCC) for pragmatic reasons. Attracting airline customers in a competitive market during this extra challenging time is inextricably linked to the strategic manipulation of linguistic resources in airline slogans. However, there has been a scarcity of contrastive analysis of rhetorical and linguistic devices used in LCCs' slogans that promote their global and local identities ascribed to airlines' culture and belief systems. Juxtaposed from McQuarrie and Mick (1996), Nilsen & Nilsen (1978, 1979), and Praba's (2017) theoretical and analytical frameworks, this study takes a corpus-based approach to analyze the rhetorical figures and linguistic devices operating in thirty (30) LCCs' slogans as represented in the three traffic conference areas which were ranked World's Best LCCs (2019-2021) by Skytrax. Findings reveal that many rhetorical figures and linguistic devices are employed in LCCs' slogans through *phonetic*, *syntactic*, and *semantic devices*. Moreover, these linguistic devices co-construct the overall rhetorical appeal of the slogans may have influenced passengers' airline choices during the pandemic. The study likewise reveals socio-cultural embeddings inferred from the airline slogans. Therefore, it can be construed that airline slogans lend awareness to sociocultural nuances framed in American, European, and Asian LCCs through the rhetorical and linguistic resources that aid in making their global and local presence and thus their identity concerns during the pandemic.

Introduction

The world witnessed the wrath of COVID-19 and how it has posed a global risk to health and economies since 2019. As of February 6, 2022, there are 414 million confirmed cases and more than 5 million fatalities related to COVID-19 (Worldometer, 2022). This pandemic has drastically affected various industries. The impact is undeniably remarkable in millions of entities (e.g., passengers, employees, companies, etc.) falling into poverty and recession. Across all industries, the aviation sector is among the hardest hit (Suau-Sanchez et al., 2020), which is seen in airline companies' predicament to survive while maintaining their credibility in a competitive market during this extra challenging time. In addition, a dwindling number of passengers has led airline companies to halt and cease almost all their operations (Sun et al., 2020a). While this difficulty has posed challenges to various legacy airline companies, it has turned all flyers across the globe into flying a low-cost carrier (henceforth, LCC) for pragmatic reasons, which put a premium on LCCs as their airline choice. Moreover, competition has been observed in how LCCs survive while attracting customers continuously in this unprecedented period.

Attracting airline customers is inextricably linked to the strategic manipulation of rhetorical devices and other linguistic resources in airline branding, trademarks, and identities that promote the credibility and resilience of LCCs. One tangible marketing strategy to attract customers with various linguistic and rhetorical devices is seen in airline companies' slogans. Airline companies that thrive on existing in unfortunate situations have to present themselves and persuade others to use their services (Laosrirattanachai, 2018) via vehicular language that carries its rhetorical appeals (e.g., emotions, reasons, and character). Such rhetorical appeals in slogans are framed through rhetorical figures and linguistic devices that may appeal to airline passengers' perception of how LCCs have made their global and local presence, and thus their identity concerns during an unprecedented time.

Hence, language in airline marketing is deemed necessary as other rhetorical resources that significantly attract customers through LCCs' slogans.

A significantly considerable number of studies have been conducted that analyze various domain-specific slogans. Such studies have been conducted to show the contrastive features of language and other rhetorical figures in different slogans; for instance, U.S. and E.U. legal protection for slogans (Petty, Leong, & Win, 2015), advertising and inspirational slogans (Fuerter-Olivera, 2001; Smirnova, 2016), corporation brand slogan (Miller & Toman, 2016), tourism slogans (Gali et al., 2017), political and advertising slogans (Keranforn-Liu, 2020), political party slogans (Koc & Ilgun, 2010). While there have been studies that employed either contrastive rhetoric and linguistic analysis or non-linguistic analysis to account for the rhetorical figures and linguistic devices used in domain-specific slogans, there are only a very few studies on how such rhetorical and linguistic devices are used in airline slogans, particularly those that are used in LCCs. The scarcity of literature on airline slogan analysis has attempted to propound the significance of slogans as a contributory feature to airline companies' marketing strategy. These slogans can be unique and appealing if constructed through linguistic perspectives, as these may help co-construct the airlines' identities. Kurniawan (2018) found that the Airline slogans used worldwide can be categorically considered phrases through a syntactic and semantic analysis of airline slogans in five continents. The majority of airlines benefit from using we are different and unique claims. Analyzing 120 food advertising slogans, Sudcharit (2015) revealed twelve figurative types: *alliteration, metonymy, hyperbole, antithesis, assonance, onomatopoeia, metaphor, pun, personification, parallelism, smile*, and *rhetorical question*. In Sudcharit's (2015) study, alliteration and parallelism appeared most frequently. Skorupa and Duboviciene's (2015) study analyzed various slogans from advertising and commercial English slogans. Their study found that figurative language should be used to attract customers. Using ideational metafunction, however, Laosrirattanachai (2018) investigated several words employed in the advertising slogans of airlines in 2016. Laosrirattanachai (2018) revealed that their slogans range from 3 to 4 to 5 words. In addition, *Airline*, *fly*, *of*, *your*, and *to* were the most frequent. In terms of ideational meta-function, it was found that participants, followed by, circumstances and processes respectively, appeared to have been demonstrated in these slogans. It is worth noting that while these studies have involved the majority of airlines as representatives of the world continent, most of the airlines involved in these studies to show airline slogans' linguistic characteristics are selected based on their presence in the global arena. These have not considered the LCCs performing well in the aviation industry. The current study argues that the LCCs, gaining either a global or local presence in the international aviation community, likewise strategically manipulate rhetorical figures and linguistic devices in their slogans, thus attracting more airline passengers to avail of their services, especially during the time of the pandemic.

Studies on advertising slogans are an interesting area for investigating how various airlines probe their marketing strategy and use a strategic manipulation of rhetorical figures and linguistic resources to gain customers and build trust and confidence in them. However, most of the studies in the existing literature have focused on the majority of airlines. As a result, contrastive analysis is scarce on how the rhetorical figures and linguistic devices operate in low-cost carriers' slogans that promote their global and local identities ascribed to airline companies' culture and belief systems. Hence, this paper explores how airline slogans are constructed and interpreted from a linguistic perspective through a contrastive analysis of their rhetorical figures and linguistic devices. Likewise, this attempts to uncover the socio-cultural attributions embedded in LCCs' slogans.

1.1 Research Questions:

1. Which rhetorical and linguistic devices are categorically and specifically employed in LCCs' slogans?
2. How do the slogans rhetorically appeal in American, European, and Asian LCCs to attract more airline passengers?
3. What socio-cultural inferences can be drawn from American, European, and Asian LCCs' slogans?

1.2 Theoretical/ Analytical Framework

The present study employs a modified framework to be used for the analysis of the LCCs' slogans. This framework juxtaposes the theoretical/ analytical frameworks used in the studies of

Miller & Toman (2016) and Kurniawan (2018). Although both studies have analyzed slogans, these were used in two domains: airline slogans and corporation brand slogans. The decision to use these two studies is seen in their theoretical underpinning, which will address the gap in analyzing rather sophisticated and dynamic socio-culturally constructed slogans of LCCs. Furthermore, to illustrate the rhetorical devices found in the study, the researcher employed an analysis of rhetorical devices, which are categorized mainly based on two schemes and tropes, based on Leigh's taxonomy (1994) cited in Monsefi & Mahadi (2017) and Laongpol (2021).

On the one hand, *Schemes* relate to syntax, word order, word omissions, insertions, letters, and sounds rather than the meaning of words. On the other hand, *Tropes* are another type of wordplay presentation that can change the ordinary meaning of words through comparison, connotation, and word choices by unusually using language. A *trope* is an artful deviation from the ordinary or principal signification, while a *scheme* is an artful deviation from the typical arrangement of words. In addition, a *trope* uses a word unusually or unexpectedly, while a *scheme* is a creative alteration in the usual order of words. In an attempt to analyze the rhetorical devices categorically, Miller and Toman (2016) yielded a large category of rhetorical devices into schemes and tropes. They analyzed specific linguistic devices categorically assigned in corporation brand slogans. Although the term linguistic devices are used by Miller and Toman (2016) rather than rhetorical figures, a specific category of linguistic devices (e.g., phonetic, syntactic, semantic) may not always capture the rhetorical construction of slogans. This is seen in the preliminary data analysis, where some rhetorical figures (in McQuarrie and Mick's terms) do not fit the specific linguistic devices. For instance, most slogans in the study have been accounted for to show the limitation of specific phonetic and syntactical devices over semantic devices. More specifically, a phonetic device *'rhyme'* differs from *'rhythm*. *Rhyming* is the practice of choosing similar-sounding words at the ends of each line, while *rhythm* is an audible pattern or effect created by introducing pauses or stressing certain words. While both can be considered phonological, these two may not share the same phonological properties but can further be explained via a syntactical device. However, to add to this dilemma, not all devices fit into the specific category of *orthographical*, *morphological*, and *syntactical devices*. For example, very little linguistic evidence in LCCs' slogans uses an *orthographic unusual or unconventional spelling*, *abbreviation*, and *word repetition*.

Analyzing the LCCs' slogans would require a categorical analysis of syntactic devices (e.g., *word, phrasal, sentential*) rather than specific, such as Praba's (2017). The current study argues that the intermarriage between rhetorical figures and linguistic devices co-constructs the rhetorical appeals of LCCs' slogans to attract more airline passengers. As a response, the present study used a modified theoretical framework based on the taxonomies of McQuarrie and Mick (1996) on rhetorical figures and Nilsen & Nilsen (1978, 1979), and Praba (2017) on linguistic analysis. This was employed in response to the first research question that accounted for the categorical and specific linguistic devices used in LCCs' slogans. Moreover, the interface between rhetorical figures and linguistic devices embedded in a slogan's rhetoric can aid in constructing its rhetorical appeals. This argues that the rhetorical figures and linguistic devices are used to show how the rhetoric of a slogan appeals to airline passengers, gives them their first impression, and eventually helps them decide whether to avail of the airline services. It can be argued that the rhetoric in airline slogans appeal to airline passengers' emotion and reasons which may be influenced by how they have perceived the airlines' values and cultures attributed to such slogans. Culture is defined as "actual practices and customs, languages, beliefs, forms of representation, and a system of formal and informal rules that tell people how to behave most of the time and enable people to make sense of their world through a certain amount of shared meanings and recognition of different meanings. Slogans are therefore seen as an embodiment of the values and cultures of the airlines, the people who work for these airlines, and probably, the people who have been attracted to these slogans as they may have found a sense of membership in the community. However, airline passengers may have different views of socio-cultural attributions to slogans, as these may vary according to their experiences and socio-lingo-cultural profile. The passengers' perceptions and impressions of airline slogans are derived from the meaning they form. Meaning lies in the power of the slogan to appeal to airline passengers' emotions and reasons. The rhetorical appeal is tied to the socio-cultural meaning perceived that may have made an impression among airline passengers. Hence, the present study likewise employs Aristotle's popular concept of rhetorical appeals, i.e., *ethos*, *pathos*, and *logos*, to analyze how airline slogans rhetorically appeal in American, European, and Asian LCCs to attract more airline passengers.

2 Methodology

The current study takes a corpus-based approach to analyze a collection of airline slogans from LCCs worldwide. A corpus-based analysis in contrastive rhetoric studies has been significant as a tool for identifying rhetorical figures and linguistic devices in various domain-specific slogans. However, it is worth noting that airline selection is established first (see Table 1). There was a stringent process in selecting the LCCs in the present study as there are thousands of LCCs worldwide.

2.1 Selection Criteria for LCCs

2.1.1 LCCs must be selected from the three traffic conference areas established by the International Air Transport Association (IATA):

Traffic Conference Area 1 (TCA1)— North and South American continents
Traffic Conference Area 2 (TCA2)— European continents
Traffic Conference Area 3 (TCA3)— Asian continents

Using the traffic conference areas is vital in clustering the LCCs based on their geographical locations on the map. The study considered these areas established by IATA over those formed by the International Civil Aviation (ICAO). This is because the ICAO does not cover the commercial matters of international airlines. At the same time, the IATA traffic conference areas were established because the traffic conferences deal with all the international air traffic matters involving passengers, cargo, and mail-in specific areas worldwide. Choosing ICAO would defeat the purpose of the study, which is to determine the linguistic devices present in the slogans of LCCs that are primarily concerned with how airlines attract more passengers and eventually avail of airline services. Hence, IATA's traffic conference areas were adopted as the main categorical criterion for selecting the LCCs.

2.1.2 LCCs must have gained an international reputation through worldwide rankings.

Different indices measure the performance of airlines as several entities rank the best airlines worldwide. Although Bazar (2019) investigated how airlines are ranked in various criteria, his study was focused on how best airline rankings are processed and indexed. It is expected that the legacy airlines that have been performing well established their reputation, as seen in the worldwide rankings produced by Airhelp – his primary source of data on worldwide ranking. In addition, the recent ranking provided by Airhelp was in 2019. However, there is no ranking by Airhelp yet, and selecting Airhelp would defeat one of the study's aims: to determine which Airline appeals to the clients during the pandemic. Hence, the present study considered the worldwide ranking provided by Skytrax annually. Since 1999, Skytrax has ranked 100 airline companies annually and evaluated the performance of the airlines based on cabin services, ground handling services, Airline, and flight products. Hence, the selection of airlines was based mainly on the worldwide rankings produced by Skytrax. Skytrax surveyed 13.42 million eligible entries that were accounted for from 2019 to 2021. This was before and during the pandemic.

Interestingly, Skytrax ranks the airlines and generates the following: World's Best Low-cost Carriers and World's Best Long Haul Low-Cost Airlines 2021 (to name a few). With the worldwide ranking provided by Skytrax, it would now be easy to determine which of the LCCs would be sampled. Following the first criterion above, the LCCs must come from the three traffic conference areas. During the preliminary sampling, it was found that in the 2021 World's Best LCCs, only three LCCs were from TCA1 and seven from TCA2. Surprisingly, ten LCCs from TCA3 emerged. Given the disproportion in the data, the researcher looked into the best LCCs in 2021, as may be represented in the TCA of IATA, which is also available on the website of Skytrax. Hence, additional seven best LCCs in TCA1 and three LCCs from TCA3 complete the thirty airlines as the representative samples of the study (See Table 1).

TCA		LCC	Origin	Slogan	No. of Words
TCA1	1	Southwest Airlines	USA	"Low fares. Nothing to hide. That's TransFarency!"	7
	2	Air Canada rouge	Canada	"Your world awaits.”	3
	3	Frontier Airlines	USA	"Low Fares Done Right".	4
	4	Spirit Airlines	USA	“Less Money. More Go”	4
	5	Sun Country Airlines	USA	“Fly at the speed of life”	6
	6	Sky Airline	Chile	"Turn around and fly"	4
	7	Easyfly	Colombia	“Easyfly makes it easy to fly”	6
	8	Gol	Brazil	"The new Gol. New times in the air."	8
	9	Viva Air	Colombia	“Fly More”	2
	10	JetSmart	Chile	"Fly SMART, fly your way.	4
TCA2	1	Vueling Airlines	Spain	"Love the way you fly"	5
	2	EasyJet	Switzerland	"This is Generation easyJet".	4
	3	Ryanair	Ireland	"Fly cheaper. The Low Fares Airline"	6
	4	Eurowings	Germany	‘ideas get wings – cha(lle)nge the future of travel’	8
	5	Norwegian	Norway	“Norwegian Airlines, the way it should be.”	7
	6	Jet2.com	UK	"Friendly Low Fares".	3
	7	Wizz Air	Hungary	"Looking ahead, only the sky is our limit."	8
	8	airBaltic	Latvia	“We Care”	2
	9	LEVEL	Spain	“It's your world.”	3
	10	Pobeda	Russia	“Rest up in Stavropol”	4
TCA3	1	AirAsia	Malaysia	“Now Everyone Can Fly”	4
	2	Scoot	Singapore	"Escape the Ordinary"	3
	3	IndiGo	India	"Go IndiGo"	2
	4	Jetstar Airways	Australia	“All day every day low fares”	6
	5	Jetstar Asia	Singapore	"All day every day low fares".	6
	6	Flynas	Saudi Arabia	"The Kingdom's First Low-Cost Airline"	5
	7	Peach	Japan	"Customers' smiles come when safety is assured"	7
	8	SpiceJet	India	Red. Hot. Spicy.	3
	9	Spring Airlines	Japan	"Don't think, Just fly!"	4
	10	Air Arabia	UAE	“Air Arabia, Pay Less Fly More.”	6
				Average No. of Words	**4.8**

Table 1. Sampled world’s best LCCs’ (2021) slogans representing TCAs

While it follows that selection of sampled airlines determined the LCCs with which the slogans of these airlines would be analyzed, a preliminary look at the number of words, which would be a point to consider in the unit of analysis, can be argued that each slogan's number of words ranges from 2 to 8. Relative to this, Laosrirattanachai (2018) revealed that the number of words in slogans ranges from 3 to 5, respectively. Essentially, the slogans of the LCCs carefully sampled from the worldwide ranking provided by Skytrax fit the required number. In addition, there were three Airline slogans with only two words; these were also included.

The data were analyzed manually using three steps: The data were matrixed to show the traffic conference areas represented by the LCCs, the country of origin, and the weighted average of words of all airlines. Likewise, each slogan was read carefully to determine the rhetorical figures and linguistic devices in all slogans. Then, using the modified framework drawn out from the taxonomies of McQuarrie and Mick (1996) on rhetorical figures and Nilsen & Nilsen (1978, 1978), Praba (2017) on linguistic analysis of LCCs' slogans, and Aristotle's rhetorical appeals, the data were read, analyzed, and labeled.

The analyzed matrix of the LCCs' slogans was then subjected to simple inter-coding reliability through the help of three inter-coders, of which agreement was reached via online consultation. The inter-coders were composed of two Ph.D. Applied Linguistics students and one Ph.D. in English degree holder; all of them are teaching in a graduate school. After the inter-coders' agreement (95%), the realization of the rhetorical figures, linguistic devices, and rhetorical appeals was discussed.

3 Results and Discussion

The first research question addresses which rhetorical and linguistic devices are categorically and specifically employed in LCCs' slogans. It can be construed that the phonetic, morphological, syntactic, and semantic devices are all used in LCCs' slogans (see Table 2). The most frequently occurring categories of linguistic devices are *syntactic* [100%] and *semantic devices* [100%], while *phonetic devices* [76.7%] appeared to be less occurring. The predominance of *morphological/ syntactical and semantic devices* is quite surprising in LCCs' slogans since this does not show a significant pattern in the previous studies (Miller & Toman, 2016; Sudcharit, 2015; Smirnova, 2016), which ranks phonologically related rhetorical devices as the highest.

3.1 Specific category of rhetorical figures and linguistic devices in LCCs slogans

However, considering the specific category of phonetic devices, the results show the unity of a phonologically related linguistic device; slogans heavily rely on *alliteration* [33.3%]. This supports the findings of Skracic et al. (2016), which revealed a high frequency of use in some slogans in yachts or boats in nautical magazines. It can be construed that the phonetic device *alliteration* aids airline slogans to be remembered easily. Recalling when a slogan alliteration is repetitive (Supphellen & Nygaardsvick, 2002; Gali et al., 2018) is more manageable. For example, 'Friendly Low Fares' (Jet2.com) shows the repetition of the first consonant sound /f/. This contrasts with Koc & Ilgon's (2010) finding, revealing *Rhyme* as the most frequent in political party slogans. While *alliteration* tops all the phonetic devices, other linguistic devices likewise occur in LCCs' slogans, such as *Assonance* [16.7%], *Rhyme* [13.3%], *Initial Plosive* [6.7%], and *Consonance* [3.3%], and *Blending* [3.3%] respectively. For instance, *assonance* is seen in "This is Generation easyJet" (EasyJet); *rhyme* in "Easyfly makes it easy to fly" (Easyfly); *initial plosive* in "The Kingdom's First Low-Cost Airline" (Saudi Arabia), and *blending* in "Low fares. Nothing to hide. That's TransFarency!" (Southwest Airlines). In terms of morphological/ syntactical devices, all LCCs' slogans' linguistic devices vary in *word*, *phrase*, and *sentential* level with a few *morphological* and *repetitional* construction occurrences. It can be deduced that LCCs' slogans are characterized as *sentential* [50.0%], followed by *phrasal* [20.0%] and *abbreviation* [13.3%]. This is followed by a few occurrences of *word/ phrase repetition* and, very rarely, one occurrence of *orthographic unusual or unconventional spelling* [3.3%] and *word* [3.3%]. On the one hand, *sentential* construction is shown in "Now Everyone Can Fly" (Air Asia), *phrasal* in "Less Money. More Go." (Spirit Airlines), and *abbreviation* in "Air Arabia, Pay Less Fly More." (Air Arabia). On the other hand, *word/phrase repetition* is also present in "Fly SMART, fly your way" (JetSmart), and occurrence of *orthographic unusual or unconventional spelling* in "Low fares. Nothing to hide. That's TransFarency!" (Southwest Airlines). While it can be construed using Praba's (2017) syntactical category would reveal that the LCCs' slogans employed *sentential* construction, Miller & Toman's (2016) categorization would show that only *abbreviation*, *word/ phrase repetition*, and *orthographic unusual or unconventional spelling* were present in the airline slogans. This concludes that the LCCs used a relatively longer slogan over phrasal slogans used in political and other advertising slogans. Regarding the semantic devices, the present study shows a significant pattern in the previous studies (Koc & Ilgon, 2010; Smirnova, 2015; Muhabat, 2015; Skracic et al., 2016; Miller & Toman, 2016; Keranforn-liu, 2020), divulging a high concentration of *metaphor* [33.3%] in slogans. It can be argued that airline slogans such as those of LCCs likewise employ *metaphor* to mention products and services indirectly efficiently (Keranforn-liu, 2020) and thus aids in attracting more airline passengers. This is followed by *personification* [26.7%] and *self-reference* [13.3%]. However, there seems to be a relative occurrence of *hyperbole* [6.7%], *antithesis* [6.7%], *paradox* [3.3%], *metonymy* [3.3%], *asyndeton* [3.3%], and *pun* [3.3%], respectively.

Variables	Frequency	Percentage
Phonetic Devices		
Alliteration	10	33.3
Assonance	5	16.7
Consonance	1	3.3
Initial Plosive	2	6.7
Blending	1	3.3
Rhyme	4	13.3
Total number containing Phonetic Devices	23.0	76.7
Morphological/ Syntactic Devices		
Orthographic unusual or unconventional spelling	1	3.3
Word/ Phrase Repetition	3	10.0
Word	1	3.3
Phrasal	6	20.0
Sentential	15	50.0
Abbreviation	4	13.3
Total number containing Syntactic Devices	30	100
Semantic Devices		
Personification	8	26.7
Metaphor	10	33.3
Self-reference	4	13.3
Paradox	1	3.3
Hyperbole	2	6.7
Metonymy	1	3.3
Pun	1	3.3
Antithesis	2	6.7
Asyndeton	1	3.3
Total number containing Semantic Devices	30	100
Rhetorical Figures		
Schemes in slogans	21	23.3
Tropes in slogans	34	37.8
Neither	12	13.3
Total number of rhetorical figures in slogans	67	74.4
Rhetorical Appeals		
Logos	9	30.0
Ethos	5	16.7
Pathos	16	53.3
Total number of Slogans containing rhetorical appeals	30	100

Table 2. Distribution of rhetorical and linguistic devices in the world's best LCCs (2021) slogans

For example, *metaphor* is seen in 'ideas get wings – cha(lle)nge the future of travel' (Eurowings), *personification* in "Fly at the speed of life" (Sun Country Airlines), and *self-reference* in "The new Gol. New times in the air." (Gol), *hyperbole* in "Looking ahead, only the sky is our limit." (Wizz Air), *antithesis* in "Less Money. More Go." (Spirit Airlines), a *paradox* in "Low Fares Done Right." (Frontier Airlines), *metonymy* in "The Kingdom's First Low-Cost Airline" (Saudi Arabia), *asyndeton* in "Red. Hot. Spicy." (SpiceJet), and *pun* in "Low fares. Nothing to hide. That's TransFarency!" (Southwest Airlines).

Analyzing the rhetorical figures, the LCCs' slogans employed more tropes [37.8%] over schemes [23.3%], while others can be neither [13.3%]. It can be deduced that tropes are more frequently used in the category of semantic devices, while schemes occur in phonetic devices. The current study reveals a similar finding that supports the high occurrence of schemes in phonetic devices and tropes in semantic devices (Miller & Toman, 2016) but shows deviance regarding the overall slogans analyzed. Miller & Toman (2016) found out that schemes are mainly used to incorporate brand slogans rather than tropes. The present study reveals that tropes are more frequently employed in airline slogans, specifically among LCCs. This suggests that LCC's slogan favors its construction through comparison, connotation, and word choices rather than word order, word omissions, letters, and sounds. Thus, LCCs' slogans are more concerned about what and how they mean than how they are arranged and sound. Therefore, it can be inferred that LCCs' slogans attract more airline passengers through semantic devices instead of phonetic devices. The airline slogan's appeal to attract more customers is related to the meaning of the rhetorical and linguistic resources rather than the sound.

3.2 Specific category of rhetorical figures and linguistic devices in LCCs slogans

Looking closely at how slogans rhetorically appeal t **Sampled world's best LCCs' (2021) slogans representing TCAs** o airline passengers that aid their decision to choose LCCs to fly, the present study reveals that LCCs' slogans heavily rely on the rhetoric of pathos [53.3] followed by logos [30.0%] and ethos [16.7%]. Concomitantly, pathos appears predominantly in all of the slogans of the LCCs as representatives of TCA1, TCA2, and TCA3. This means that slogans from the Americas, Europe, and Asia LCCs emphasize pathos to appeal to airline passengers' emotions. This rhetorical appeal is gleaned more predominantly from TCA2, suggesting that European LCCs are more drawn from making their slogans appeal to airline passengers' emotions over logic and authority presence. While this also occurs predominantly in TCA1, suggesting the same rhetorical appeals from the American LCCs, its use in TCA3 suggests that Asian LCCs' slogans appeal to airline passengers' emotions but remain intuitive to logic and authority presence. Therefore, it can be construed that American, European, and Asian airlines construct slogans where rhetorical appeals operate to attract more customers to avail of their services.

The present study results suggest that linguistic and rhetorical devices are frequently employed in airline slogans, precisely that of LCCs. Although the LCCs' slogans' number of words ranges from 2 to 8, a closer look at the syntactic devices reveals that slogans at the sentential level may appeal rhetorically to airline passengers. It can be argued that LCCs' slogans may not favor the sound that creates an impression among the airline passengers but can appeal rhetorically to passengers' emotions through the slogans constructed in length. This is seen in how the average number of words in LCCs' slogans in TCA2 [5.0] was employed more than in TCA1 [4.8] and in TCA3 [4.6]. Although slogans need to be simple, these seem to be moderately complex (Miller & Toman, 2016) in using linguistic devices. Using these linguistic devices can influence airline passengers' memory (Nilsen & Nilsen, 1978) which may generate positive affective responses (McQuarrie and Mick 1996). These linguistic units are contained in phonetic, syntactic, and semantic devices. The most frequently occurring category of phonetic devices shows that American and Asian LCCs commonly employ alliteration and Rhyme in their slogans. However, the most common syntactic category of linguistic devices is sentential, which is heavily seen in European LCCs'. The commonly utilized semantic device appears to manifest predominantly in European LCCs' slogans.

However, the use of tropes in airline slogans reveals a different pattern from the previous studies on domain-specific slogans. It can be deduced that tropes are more frequently used in the category of semantic devices, while schemes occur in phonetic devices. The current study reveals a similar finding that supports the high occurrence of schemes in phonetic devices and tropes in semantic devices (Miller & Toman, 2016) but shows deviance regarding the overall slogans analyzed. Therefore, it can be concluded that tropes operate more than schemes in LCCs' slogans. Specifically, tropes are heavily used from TCA1, suggesting that American LCCs are more concerned about the meaning of the slogans than the sound and construction. Therefore, it can be inferred that LCCs' slogans attract more airline passengers through semantic devices instead of phonetic devices. Finally, the rhetoric of pathos appears to be gleaned from the LCCs' slogans of European airlines, which prioritized an appeal to emotion over logic and authority presence. Therefore, it can be construed that American, European, and Asian airlines construct slogans where rhetorical appeals operate to attract more customers to avail of their services.

Looking at how rhetorical figures, linguistic devices, and rhetorical appeals operate in airline slogans, it can be construed that LCCs in American, European, and Asian LCCs frame their slogans by establishing a strategic manipulation of linguistic recourses that aim to have made their global and local presence and thus their identity concerns during the pandemic. While airline companies thrive on existing in unfortunate situations, they still have to present themselves and persuade others to use their services (Laosrirattanachai, 2018) via a vehicular language that carries its rhetorical appeals (e.g., emotions, reasons, and character) and their identities during the pandemic. Therefore, LCCs' slogans observed relatively more complex linguistic units than simple ones. Although it can be sensitized that the simpler a slogan is, the easier it is to appeal to airline passengers, I argue that the complexity of a slogan creates a strong impression among airline passengers when focused on its meaning. This can be inferred from how European LCCs' slogans are predominantly framed in length but still manage to have made their global and local presence, and thus identity concerns during an unprecedented time. The same can be observed in American and Asian LCCs slogans that thrive on making their presence and creating identities through a relatively lengthy linguistic pattern.

3.2 Socio-cultural inferences from American, European, and Asian LCCs' slogans

Indeed, rhetorical figures and linguistic devices co-occur in LCCs' slogans to create a rhetoric that would appeal to passengers' airline choices. It is undeniably argued that the LCCs' slogans analyzed in the present study have revealed contrastive rhetoric that would create an impression among airline passengers. It has been seen that American LCCs pay more attention to the semantic aspect of slogans than the phonetic features. On the other hand, while European LCCs tend to focus on the length of the slogans that would affect the overall impression among airline passengers, Asian LCCs would emphasize a relatively lengthy slogan to appeal to the same impression while maintaining a collective identity and authority presence. It is then worth mentioning that the slogans of LCCs have socio-cultural embeddings that may contribute to attracting airline passengers. For example, the slogans of the Asian LCCs have been observed to show how they project a collective identity and authority presence, as seen in the slogan of Flynas (Arabia), The Kingdom's First Low-Cost Airline. This slogan arguably protrudes that Arabia airline enjoins the community it serves to project solidarity, unity, and collective identity. AirAsia's "Now Everyone Can Fly" slogan can illustrate a similar observation. Malaysia has projected Air Asia's slogan to embed a socio-cultural feature of Asian collectivism among airline passengers. This sociocultural inference can be interpreted by Boiger et al. (2012), concluding that people in East- Asian cultural contexts emphasize adjusting themselves to fit in with (the role requirements of) their social environments (Morling & Evered, 2006; Morling, Kitayama, & Miyamoto, 2002; Weisz, Rothbaum, & Blackburn, 1984, cited in Boiger et al., 2012).

Moreover, Confucian values—like respect for authority and desire for harmony—are highly respected in Asian societies. The concept of mien tsu, which stands for prestige, is a function of social status and constant pressure to live up to the community's expectations. In Asian societies, individuals consume commodities that measure their social class or enhance their status. Asian individuals feel a strong need to improve their position in society. It can be inferred that indexicality in airline slogans can be construed from Asian LCCs' slogans which symbolize solidarity, unity, and collectivism. However, some socio-cultural inferences can also be gleaned from American and European LCCs' slogans. For example, Southwest Airlines (USA) slogan, "Low fares. Nothing to hide. That's TransFarency!", Spirit Airline's (USA) "Less Money. More Go", Sun Country Airline's (USA) "Fly at the speed of life," and Frontier Airline's (USA) Low Fares Done Right" have favored the use of semantic devices that create an impactful meaning among airline passengers. It can be argued that the American LCCs' slogans may have projected their identity that fosters a unique concern as they make their local and global presence. This can be explained in Boiger et al.'s (2012) argument that American contexts tend to construct action to influence their environment to make the environment fit their concerns.

On the other hand, the European LCCs' slogans have also manifested socio-cultural inferences that attract more airline passengers. Most European LCCs pay attention to the complex morphology of slogans, as evident in lengthy construction, rather than the semantic and phonetic aspects. This infers that European LCCs project a culture that identifies them. This is seen in Eurowings' (Germany) slogan 'ideas get wings – cha(lle)nge the future of travel' and Norwegian (Norway) "Norwegian Airlines, the way it should be," which projects selfhood characteristic of Europeans while maintaining their commitment to others as evident in their egalitarian values. This can be referred back to numerous attempts of European LCCs to include a self-reference strategy in constructing their slogans. A distinctive European value can be inferred from the excessive use of self-reference, which endorses selfhood, but remains committed to serving others. It can be deduced that European LCCs strongly support Harmon – egalitarianism rather than hierarchy. Their LCCs ' slogans show commitment to others and egalitarianism rather than individualism.

4 Conclusion and Recommendation

This study analyzed the rhetoric of thirty LCC slogans, carefully selected from the three traffic conference areas established by IATA. The world’s best LCCs of the year 2021 named by Skytrax include LCCs from TCA1: Southwest Airlines (USA), Air Canada rouge (Canada), Frontier Airlines (USA), Spirit Airlines (USA), Sun Country Airlines (USA), Sky Airline (Chile), Easyfly (Colombia), Gol (Brazil), Viva Air (Colombia), JetSmart (Chile). From TCA2, World’s Best LCCs include Vueling Airlines (Spain), EasyJet (Switzerland), Ryanair (Ireland), Eurowings (Germany), Norwegian (Norway), Jet2.com (U.K.), Wizz Air (Hungary), airBaltic (Latvia), LEVEL (Spain), Pobeda (Russia). Finally, from TCA3, Skytrax named World’s Best LCCs, which include AirAsia (Malaysia), Scoot (Singapore), IndiGo (India), Jetstar Airways (Australia), Jetstar Asia (Singapore), Flynas (Saudi Arabia), Peach (Japan), SpiceJet (India), Spring Airlines (Japan), and Air Arabia (UAE).

This study revealed a corpus-based analysis that many rhetorical figures and linguistic devices are employed in LCCs’ slogans through phonetic, syntactic, and semantic devices. In addition, such linguistic devices co-construct the overall rhetorical appeal of the slogans that may have influenced passengers’ airline choices during the pandemic. Furthermore, the study argued that rhetorical figures, linguistic devices, and rhetorical appeals are features of airline slogans in American, European, and Asian LCCs. Finally, the present study reveals snippets of socio-cultural embeddings inferred from the airline slogans as evident in American, European, and Asian LCCs. It has been concluded that American LCCs’ slogans project individualism through semantic devices. And while Asian LCCs’ slogans index collectivism to show solidarity and unity, European LCCs have conducted selfhood as characterized by the use of self-reference in their slogans but maintain a commitment to others as a manifestation of egalitarian values among Europeans. Although the study has empirically investigated the linguistic features of airline slogans, specifically among LCCs that have made a strong presence during the pandemic, further research can be explored, including other prestigious airline companies that rank the world’s best airlines since only Skytrax was chosen as the primary database as it is considered the most relevant data source for this study. In addition, other entities such as Star Alliance, Sky Team, and One World groups may be incorporated into future studies.

Acknowledgments

I thank Dr. Alain Razalan, Dr. Helen Trindad, Prof. Myla Magayanes, Prof. Grace Cerdan, and Jairus Duque for their assistance in various ways.

References

Ozlem Deniz Basar. 2019. Comparison of world airline rankings with different criteria: Best airline ranking and EVAMIX method rank. *American Research Journal of Humanities Social Science, 2*(3): 38-46.

Michael Boiger, Batja Mesquita, Annie Y. Tsai, and Hazel Markus. 2012. Influencing and adjusting in daily emotional situations: A comparison of European and Asian American action styles, *Cognition & Emotion*, 26(2):332-340.

Coronavirus cases: Worldometer. (n.d.). Retrieved February 6, 2020, from https://www.worldometers.info/coronavirus/

Pedro A. Fuertes-Olivera, Marisol Velasco-Sacristan, Ascension Arribas-Bano, and Eva Samaniego-Fernandez. 2001. Persuasion and advertising English: Metadiscourse in slogans and headlines. *Journal of Pragmatics,* 33:1291-1307.

Nuria Gali, Raquel Camprubi, Jose A. Donaire. 2017. Analysing tourism slogans in top tourism destinations, *Journal of Destination Marketing & Management,* 6(3):243-251.

International Air Transport Organization. *Provisions for the Conduct of the IATA Traffic Conferences.* https://www.iata.org/

Loic Keranforn-Liu. 2020. *The linguistic characteristics of political and advertising slogans: A contrastive analysis.* [Doctoral dissertation, University of Toulouse] Department of the English World Studies. Repository. https://dante.univ-tlse2.fr/access/files/original/

Erdogan Koc and Ayse Ilgun. 2010. An investigation into the discourse of political marketing communications in Turkey: The use of rhetorical figures in political party slogans, *Journal of Political Marketing,* 9(3):207–224.

Iwan Kurniawan. 2018. The language of airline slogans: A linguistics analysis. English Education: *Jurnal Tadris Bahasa Inggris,* 11(1):59–81.

Jitsuda Laongpol. 2021. A contrastive study on rhetoric in COVID-19-related news headlines from native and non-native English online newspapers, *The Southeast Asian Journal of English Language Studies,* 27(1): 47-61.

Piyapong Laosrirattanachai. 2018. An Analysis of Slogans of Airline Business Using Ideational Metafunction. *Humanities Journal,* 25(1):316-343.

Edward F. McQuarrie and Mick David Glen. 1996. Figures of rhetoric in advertising language. *Journal of Consumer Research,* 22(2):424-438.

Darryl W. Miller and Marshall Toman. 2016) An analysis of rhetorical figures and other linguistic devices in corporation brand slogans. *Journal of Marketing Communications,* 22(5):474-493.

Roya Monsefi, Tengku Sepora Tengku Mahadi. 2016. Wordplay in English online news headlines. *Advances in Language and Literary Studies,* 7(2):68-75.

Fakharh Muhabat, Mehvish Noor, and Mubashir Iqbal. 2015. Advertisement of School Slogan: Semantic Analysis. *European Academic Research*, 3(1):419-433.

Don L.F. Nilsen. 1979. "*Language Play in Advertising: Linguistic Invention in Product Naming*." In Georgetown University Round Table on Language and Linguistics 1979, edited by James E. Alatis and G. Ricahrd Tucker, 137–143. Washington, DC: Georgetown University Press.

Don L.F. Nilsen and Alleen Pace Nilsen. 1978. *Language Play: An Introduction to Linguistics*. Rowley, MA: Newbury House Pub.

Ross D. Petty, Susanna H.S. Leong, and May O. Lwin. 2015. Slogans: U.S. and E.U. legal protection for slogans that identify and promote the brand. *International Journal of Advertising,* 29(3):473-500.

U Praba. 2017.. Grammar Analysis. Pamulang. Penerbit UT.

Tomislav Skračić and Petar Kosović. 2016. Linguistics Analysis of English Advertising Slogan in Yachting. Transactions on Maritime Science, 5(1);40-47.

Pavel Skorupa and Tatjana Dubovičienė. 2015. Linguistic characteristics of commercial and social advertising slogans. *Coactivity: Philology, Educology,* 23(2):108-118.

Skytrax. 2021. World's best low-cost carriers 2021. https://www.worldairlineawards.com/worlds-best-low-cost-airlines-2022/

Tatjana Smirnova. 2016. Sound of a slogan: appealing to audiences in the global market. *Procedia – Social and Behavioral Sciences,* 236:125-130.

Pere Suau-Sanchez, Augusto Voltes-Dorta, and Natalia Cuguero-Escofet. 2020. An early assessment of the impact of COVID-19 on air transport: Just another crisis or the end of aviation as we know it? *Journal of Transport Geography*, 86: 102749.

P. Sudcharit. 2015. Figurative language in food advertising slogans (A Special Study Report for the Degree of Master of Arts Program in English for Professional and International Communication). King Mongkut's University of Technology Thonburi, Bangkok.

Xiaoqian Sun, Sebastaian Wandelt, and Anming Zhang. 2020. How did COVID-19 impact air transportation? A first peek through the lens of complex networks. *Journal of Air Transport Management,* 89:101928.

Cross-strait Variations on Two Near-synonymous Loanwords *xie2shang1* and *tan2pan4*: A Corpus-based Comparative Study

Yueyue Huang[1,2]
[1] Guangzhou Xinhua University, 19 Huamei Road, Tianhe District, Guangzhou, Guangdong, China;
[2] The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China
hailey.huang@connect.polyu.hk

Chu-Ren Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China
churen.huang@polyu.edu.hk

Abstract

This study attempts to investigate cross-strait variations on two typical synonymous loanwords in Chinese, i.e. 协商(*xie2shang1*) and 谈判 (*tan2pan4*), drawn on MARVS theory[1]. Through a comparative analysis, the study found some distributional, eventual, and contextual similarities and differences across Taiwan and Mainland Mandarin. Compared with the underused *tan2pan4*, *xie2shang1* is significantly overused in Taiwan Mandarin and vice versa in Mainland Mandarin. Additionally, though both words can refer to an inchoative process in Mainland and Taiwan Mandarin, the starting point for *xie2shang1* in Mainland Mandarin is somewhat blurring compared with the usage in Taiwan Mandarin. Further on, in Taiwan Mandarin, *tan2pan4* can be used in economic and diplomatic contexts, while *xie2shang1* is used almost exclusively in political contexts. In Mainland Mandarin, however, the two words can be used in a hybrid manner within political contexts; moreover, *tan2pan4* is prominently used in diplomatic contexts with less reference to economic activities, while *xie2sahng1* can be found in both political and legal contexts, emphasizing a role of mediation.

1 Introduction

Research into near-synonyms in Mandarin Chinese, particularly verbal ones, has attracted scholarly attention in recent years. Exploration of semantic differences of near-synonyms can contribute to our knowledge of the Chinese language. As "syntactic behaviours of verbs are semantically determined" (Chief et al., 2000, p. 57), a comparison of the syntactic information of synonymous verbs thus can effectively reveal semantic differences between the verbs (Chief et al., 2000).

Nonetheless, the syntactic information is sometimes intricated to catch and can be confusing to scholars of interest for lacking a clear representation of semantic clues hidden in syntactic information. In response to it, the Module-Attribute Representation of Verbal Semantics (MARVS) theory was then proposed to construct Chinese verbal semantics better. The theory was based on the premise that lexical semantic representation is the grammaticalization of conceptual information, i.e., they can be linked to grammatical structure with conceptual motivation and be attested by representational clues (Chung & Ahrens, 2008; Huang et al., 2000; Huang & Hsieh, 2015). Representational clues include collocation, argument section constraints, distributional patterns along with other elements that can be attested by corpus evidence (Huang & Hsieh, 2015).

MARVS theory denotes that verbal semantics can be differentiated based on eventive information, which is comprised of event modules and role modules, both bearing its internal attributes (see Figure 1). There are five 'atomic event

[1] The hereafter number "1, 2, 3, 4" followed after Pinyin corresponds to the four Chinese tone mark [ˉ ˊ ˇ ˋ].

structures' in the event modules, including (Huang et al., 2000, p.26):

(1) **.** Boundary: it can be identified with a temporal point, and that must be regarded as a whole.

(2) **/** Punctuality: a single occurrence of an activity that cannot be measured by duration.

(3) ///// Process: an activity that has a time course.

(4) _____State: a homogeneous module in which the concept of temporal duration is irrelevant.

(5) ^^^^^ Stage: a module which consists of iterative sub-events.

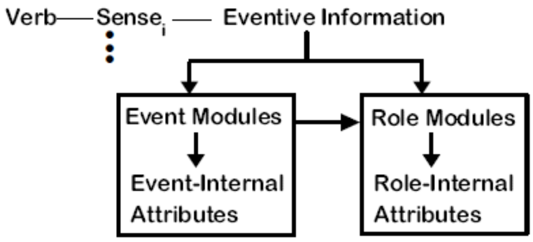


Figure 1. Theoretical framework of MARVS (Huang et al., 2000)

The role modules include the focused roles (participants) of an event and the role-internal attributes, the latter being "the semantic properties of the participants, such as [sentience], [volition], and so forth" (Huang & Hsieh, 2015)

Under the framework of MARVS, a considerable number of studies have been carried out on Chinese verbal near-synonyms (Ahrens et al., 2003; Chung & Ahrens, 2008; Tao, 2021; Wang & Huang, 2018). Further studies on MARVS expanded our understanding of aspects related to Chinese verbal near-synonyms, including their spatial-temporal attributes (Liang & Huang, 2021) and mental states (Tao, 2021),

Additionally, as cross-strait variation has also been observed by scholars (Hong & Huang, 2008; Hung et al., 2007), a more specific focus on issues related to such variations using MARVS has been made to enrich our views on the diversity within the Chinese language. A few issues have been touched upon, such as power relation (Wang & Huang, 2018) and viewpoint foci differentiation (Wang & Huang, 2021). Nonetheless, fine-grained semantic relations on cross-strait variations are still in need of further investigation.

This study then attempts to contribute to our understanding of cross-strait variations on synonymous loanwords in Chinese, taking two typical loanwords as an entry point, i.e. 协商(*xie2shang1*) and 谈判 (*tan2pan4*). MARVS theory was adopted as the analytical framework to construct a fuller picture of the semantic variations of the two words.

2 The present study

2.1 Loan words in Chinese

Chinese orthographical or syntactic system has not been static. It has undergone generations of evolutions with both in-group and out-group momentum. One typical out-group linguistic influence comes from cross-cultural contacts, hence ubiquitous loanwords observed in Chinese language systems. Loanwords could come from English (Kim, 2018), Japanese (Shi, 2020), and many other languages (e.g. Russian) that might have come into contact with China throughout history.

In particular, a large number of political or social terms were introduced in the Chinese language from the late 1800s to the 1900s, a time of social turmoil (Gunn, 1991; Masini, 1993). The social terms 谈判 (*tan2pan4*) and 协商 (*xie2shang1*), listed in official Chinse loanwords dictionaries (Shi, 2019, p.1108+1253), were introduced into modern Chinese roughly at such time. The former was considered to be translated into Chinese through relay translation (English to Japanese to Chinese) in the period of 1840~1920s (from the Opium War of 1840 to before the Anti-Japanese War) (Shi, 2019, p.1108; Chen, 2014); while the later regarded as a merged, interchangeable term with both Japanese and Chinese word 协议 (*xie2yi1*, agreement), which has been influenced from Japanese in the late 1800s as well (Shi, 2019, p.1253), and eventually formed what we observed in modern usage as it might later have mixed influence from a Latin word, *deliberationem* (Wang & Zhang, 2010).

To confirm what was found in the literature, a balanced one billion-bytes CCL corpus[2] and Google N-gram Viewer[3] were chosen to examine the traces of their usage in history. The CCL corpus (Center for Chinese Linguistics PKU) consists of 581,794,456 modern Chinese characters and 201,668,719 ancient Chinese characters (Zhan et al., 2019). Google N-gram is an online search engine that charts the frequencies of search in printed books between 1500 and 2019 in text corpora collected by Google in eight languages, including Chinese (Michel et al., 2011).

[2] http://ccl.pku.edu.cn:8080/ccl_corpus/

[3] https://books.google.com/ngrams

The two words were first searched in the ancient Chinese subcorpus of CCL. The term *xie2shang1* generated 63 hits, with the earliest usage traced back to the late Qing Dynasty. And the search for *tan2pan4* retrieved 76 hits with the earliest usage found in the period of the Republic of China. Selected examples in CCL are listed as follows:

Example 1: 丁亥，命班第赴金川军营**协商**军务。(二十五史\清史稿)

Pinyin: *ding1hai4, ming4 ban1di4 ban1di4 fu4 jin1chuan1 jun1ying2* ***xie2shang1*** *jun1wu4.*

Translation: In the year of Dinghai, Bandi was ordered to go to the Jinchuan military camp to **consult on** military affairs.

Example 2: 要求贵国即予同意，……速开正式**谈判**……（民国\小说\民国演义）

Pinyin: *yao1qiu2 gui4guo2 ji2yu3 tong2yi4, … su4 kai1 zheng4shi4* ***tan2pan4****…*

Translation: We hope that you can agree to open a formal **negotiation** soon.

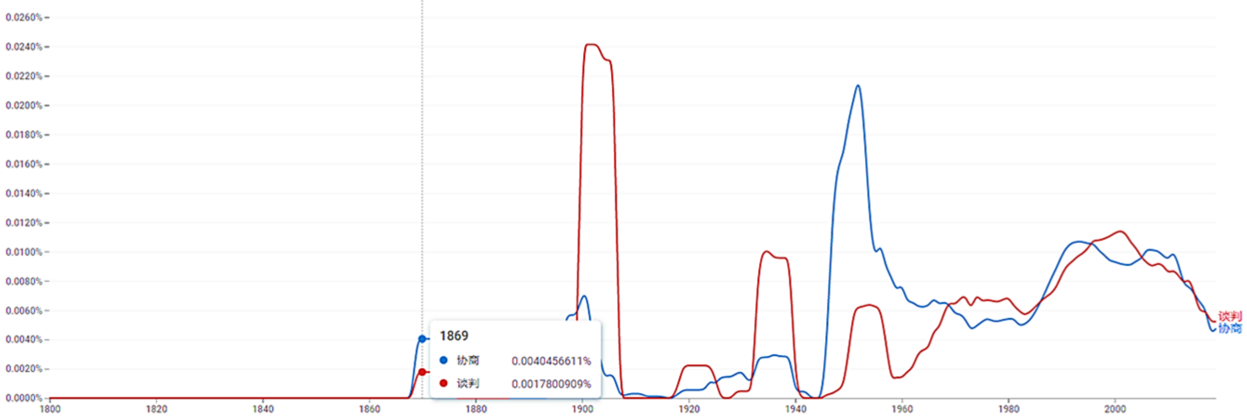


Figure 2 Google N-gram results for *xie2shang1* and *tai2pan4*

To double-check the result, the authors also generated Google N-grams of the two words, as shown in Figure 2. Both words appeared roughly at the same time in the 1870s, which fall into the generalized findings in the previous literature. Subsequent searches through Google Books provided examples in Japanese texts, but, through screening, the earlier attested examples for the two words in Chinese were found in 清議報 (*Ching i Po*, Qingyi Newspaper), published by the Royalists led by Liang Qichao in Japan in 1900. Cross-reference of all sources, including literature, CCL corpus, Google Books, and Google N-gram results, suggested that the two words came into modern Chinese around the second half of the 19th century.

2.2 Research Questions

As the two words (i.e., *xie2shang1* and *tan2pan4*) were established as loan words, the present study was then valid to proceed. The study attempts to analyze cross-strait variations of these two synonymous loan words using MARVS theory. More precisely, it attempts to answer the following research questions:

RQ1: What are the distributional differences of *xie2shang1* and *tan2pan4* between Mainland Mandarin and Taiwan Mandarin?

RQ2: What are the event representations of *xie2shang1* and *tan2pan4* in Mainland Mandarin and Taiwan Mandarin?

RQ3: What are the role representations of *xie2shang1* and *tan2pan4* used in Mainland Mandarin and Taiwan Mandarin?

2.3 Research Method

The Gigaword Corpus (CWS)[4] and its two subcorpora were chosen for this study — Gigaword_XIN (XIN) and Gigaword_CNA (CNA) via Chinese Word Sketch (Hong et al., 2006; Ma et al., 2006). The former (XIN) was compiled by news texts from Xinhua News Agency of Beijing (382,881,000 tokens), and the latter (CNA) by news from the Central News Agency of Taiwan (735,499,000 tokens) (Huang & Wang, 2020).

In line with the MARVS-based lexical semantics methodology proposed in the studies of (Chung & Ahrens, 2008; Huang et al., 2000), the

[4] https://wordsketch.ling.sinica.edu.tw/

present study will follow the research process to answer the above questions as stated:

First, to establish the near-synonymous relationship of the two words by analyzing their senses based on the meanings in the Chinese WordNet and examining the examples in the main corpus.

Second, to examine the distributional patterns of the two words in Mainland and Taiwan Mandarin.

Third, to analyze their collocations (e.g., modifier/modified, propositional phrases) to construct their event representations.

Forth, to analyze the agent-goal/subject-object relationship and the role internal attributes of the two words to construct their role representations.

3 Findings and Discussions

3.1 Establishing near-synonyms

It is somewhat tricky to establish the near-synonymous semantic relations between *xie2shang1* and *tan2pan4*. In Chinese WordNet (CWN)[5], compiled by the Institute of Linguistics, Academia Sinica, *xie2shang1* means "reaching a consensus through discussion among disagreeing parties" (意见不同的几方一起讨论以取得各方都能接受的结论; *yi2jian4 bu4tong2 de ji1fang1 yi4qi3 tao3lun4 yi3 qu3de2 ge4fang1 dou1 neng2 jie1shou4 de jie2lun4*). It can serve both as an intransitive VERB and a NOUN. The definition provided by CWN suggests that the word is an action involving more than two parties, with the potential of a mediator, and it aims for a win-win, optimizing outcome for all.

In the case of *tan2pan4*, CWN does not manage to compose its corresponding meaning description, yet is able to indicate this word is a combination of two morphemes *tan2*, meaning "to exchange information in a verbal manner" (用言语交换讯息; *yong4 yu3yan2 jiao1huan4 xun4xi1*) and *pan4*, meaning "to make a conclusion about the subsequent events based on certain criteria" (根据特定标准对后述事件做出结论;*gen1jun4 te4ding4 biao1zhun3 dui4 hou4shu4 shi4jian4 zuo4chu1 jie2lun4*). In this sense, the word can consequently mean "to make a conclusion about certain events through conversation or discussion". It thus refers to an action that might typically involve two parties, optimizing the outcome for either party (a zero-sum game).

Examination for concordances in CWS confirmed certain interchangeability of the two words. For example:

1) ……将继续和市府及业者**协商**，使运价更合理。

Pinyin: *...jiang1 ji4xu4 he2 shi4fu3 ji2 ye4zhe3* ***xie2shang1****, shi3 yun4jia4 geng4 he2li3.*

Translation: ...will continue **negotiating** with the city government and the industry to make the tariff more reasonable.

2) 不断完善共产党领导的多党合作和政治**协商**制度…

Pinyin: *bu4duan4 wan2shan4 gong4chan3dang3 ling3dao3 de duo1dang3 he2zuo4 he2 zheng4zhi4* ***xie2shang1*** *zhi4du4…*

Translation: …continue to improve the system of multi-party cooperation and political **consultation** led by the Communist Party…

3) 如果美国不与伊拉克**谈判**，伊拉克就不会退出科威特。

Pinyin*: ru2guo3 mei3guo2 bu4 yu3 yi1la1ke4* ***tan2pan4****, yi1la1ke4 jiu4 bu4hui4 tui4chu1 ke1wei1te4.*

Translation: Iraq will not withdraw from Kuwait if the US does not **negotiate** with Iraq.

Since the near-synonymous relationship of the two words is established, it is then to analyze the cross-strait variations in terms of their distributions, event modules and role modules.

3.2 Distributional Variations

To examine their cross-strait variations, we searched two node words in both subcorpora CNA and XIN, and the result can be seen in Table 1. As the two subcorpora contain a disproportionate number of corpus data, a log-likelihood formula[6] was run to compare the frequency distribution across the two sub-corpora.

Table 1 Distributional variations of *xie2shang1* and *tan2pan4* in CNA and XIN

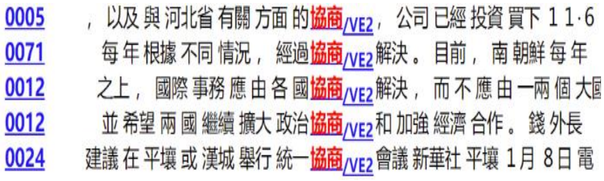
	Freq. in CNA	**Freq. in XIN**	**Log-Likelihood**	**sig.**
谈判	111,619	67,301	894.52	0.000 *** -
	Freq. in CWS			180,550
协商	91,998	20,215	14604.35	0.000 *** +
	Freq. in CWS			112,649

[5] https://cwn.ling.sinica.edu.tw/

[6] https://ucrel.lancs.ac.uk/llwizard.html

The results suggest that, compared with the significantly underused tan2pan4, xie2shang1 is significantly overused in Taiwan and vice versa in Mainland China.

Additionally, CWS PoS results indicate that *tan2pan4* can serve as both activity transitive verbs (VC2) and as a common noun (Na), while *xie2shang1* can only be activity verbs with a sentential object (VE2). Both actions indicate the dual participation of agent as subject and goal as the object. The grammatical category of *xie2shang1* found in Gigaword Corpus seems to show a discrepancy with what was concluded in CWN, issued by a prestigious linguistic institute in Taiwan, which suggests it can be both a verb and a noun. The concordance search in XIN was then run to confirm whether there is a valid discrepancy in the grammatical categories of xie2shang1 between Taiwan and Mainland Mandarin. It was found that the word, though sometimes tagged as VE2, could still present deverbal features in context, such as in No.0005 in Figure 3. It implies that the word might be in a blurring grammatical position between verbal and deverbal elements in Mainland Mandarin.



0005 ，以及與河北省有關方面的協商/VE2，公司已經投資買下 1 1·6
0071 每年根據不同情況，經過協商/VE2解決。目前，南朝鮮每年
0012 之上，國際事務應由各國協商/VE2解決，而不應由一兩個大國
0012 並希望兩國繼續擴大政治協商/VE2和加強經濟合作。錢外長
0024 建議在平壤或漢城舉行統一協商/VE2會議新華社平壤 1月 8日電

Figure 3 Concordances of *xie2shang1*(協商/协商) in XIN

One possible explanation for this might be that the overuse of *xie2shang1* in Taiwan Mandarin broadens the scope of its grammatical categories, while underusage of *xie2shang1* in Mainland Mandarin may still experience an ongoing and changing process for the word's grammatical attributes. This further points to a subtle linguistic change that might occur with a disproportionate usage of the same language in different geographical locations.

3.3 Event Representation

WordSketch for each word was run in both XIN and CNA to examine their semantic variations in collocation. As each word yielded different occurrences in two subcorpora, the nominalized frequency was provided to establish a comparable baseline and multiplied by 10,000 for the purpose of clearer presentation.

All collocated results were examined to construct the event modules of the words, including the returned lists of their possessors, possessions, modifiers, the modified and propositional phrases. Typical collocations indicating one of the 'atomic event structures' were extracted, and the top four frequent ones were shown in Table 2 and Table 3. Table 2 included one more collocated word, "中" (*zhong1;* middle), as it typically occurs as part of the syntactic structure "正在……中……" (*zheng4zai1…zhong1*; being in the middle of…).

Table 2 Collocation of *tan2pan4* across CNA and XIN

谈判	collo-cation	F.	NF*	T-score	MI
CNA =111,619	迄	12	**1.08**	3.46	14.15
	今	12	**1.08**	3.46	12.54
	开始	245	21.95	15.65	11.24
	正在	153	13.71	12.37	12.46
	中	26	2.33	5.01	5.77
	进展	137	12.27	11.70	14.08
XIN =67,301	迄	12	**1.78**	3.46	13.91
	今	12	**1.78**	3.46	13.18
	开始	385	57.21	19.61	10.46
	正在	74	11.00	8.60	11.13
	中	30	4.46	5.43	6.83
	过程	212	31.50	14.56	12.43

**NF: Nominalized frequency=Frequency/10,000*

Table 2 shows that both in Taiwan and Mainland Mandarin, the event structures of *tan2pan4* point to both boundary and process. For its boundary structure, the word *tan2pan4* has a clear starting point, as *kai2shi3* (开始; start to) can be used with *tan2pan4* (MI=11.24/10.46); however, its ending point is not very clear as the only one collocated propositional phrase *qi4jin1* (迄今; so far) only occur roughly ONCE in 10,000 times either in Taiwan or Mainland Mandarin. Furthermore, the phrase 'so far' tends to point to a middle point during such an event process. Its process indicator is rather prominent as it can both be collocated with process indicators, such as 正在谈判中 (*zheng4zai4 tan2pan4 zhong1*; is under negotiation…), 谈判进展 (*tan2pan4 jin4zhan3*; negotiation progress) or 谈判过程 (*tan2pan4 guo4cheng2*; the process of negotiation). Thus *tan2pan4* can be considered an inchoative process in Taiwan and Mainland mandarin (• /////).

The collocations with the word *xie2shang1* similarly indicate that this word can refer to a process event with a starting point, yet no ending point (see Table 3). Typical process indicators other than 正在(*zheng4zai4*), which was mentioned above, also include 继续 (*ji4xu4*; continue to...), 持续 (*chi2xu4*, continue to…), etc.. Additionally, though both Taiwan and Mainland usage of *xie2shang1* have the indicator (*kai2shi3*) for its starting point of a boundary, comparing the nominalized frequency of *kai2shi3* suggests that the starting point in its Taiwan usage is prominently clearer than that in its Mainland Mandarin (NF in CNA=41.85 > NF in XIN=8.41). So even though the word can be an inchoative process in Taiwan Mandarin (• */////*) and Mainland Mandarin, the starting boundary in Mainland Mandarin (• */////*) is somewhat blurring, comparatively speaking.

Table 3 Collocation of *xie2shang1* across CNA and XIN

协商	collocation	F.	NF*	T-score	MI
CNA =91,998	继续	911	99.02	30.17	11.15
	持续	68	7.39	8.24	11.87
	正在	75	8.15	8.66	12.46
	开始	385	**41.85**	19.61	10.46
XIN =20,215	继续	55	27.21	7.41	10.89
	坚持	31	15.34	5.57	11.70
	正在	57	28.2	7.55	11.13
	开始	17	**8.41**	4.12	10.46

**NF: Nominalized frequency=Frequency/10,000*

It is also worth mentioning that both words in XIN and CNA have disposal inherent event attributes, as both can be collocated with *ba*, for example:

(1) ……一直反对把贸易谈判跟其他政治问题扯在一起。

Pinyin: *yi1zhi2 fan3dui4 ba3 mao4yi4 tan2pan4 gen1 qita1 zheng4zhi4 wen4ti2 che3 zai4 yi1qi3…*

Translation: …has always been opposed to tying trade talks with other political issues.

(2) 把总预算案协商，是为政治角力工具。

Pinyin: *ba3 zong3 yu4suan4 an4 xie2shang1, shi4wei2 zheng4zhi4 jue2li4 gong1ju4.*

Translation: … to take the general budget consultation as a tool of political wrestling.

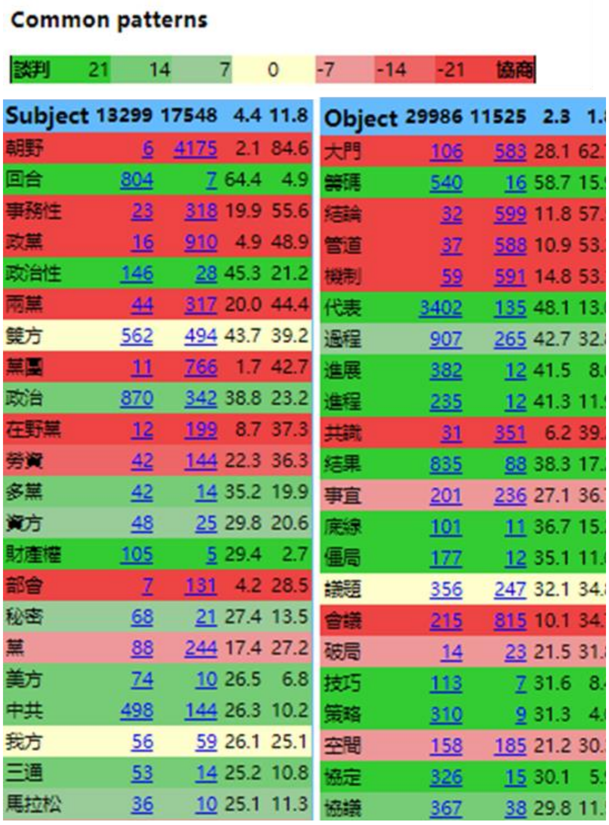
Findings in this session point to little distinctions of the words across Mainland and Taiwan Mandarin. It is understandable as the two words were introduced to the Chinese language only roughly over a century ago, and they have not undergone many changes throughout time. Nonetheless, their role representations might well exhibit quite diverged contextual variations as the words experienced social turmoil since their introduction. The following session will compare their role representations in XIN and CNA.

3.4 Role Representation

Common patterns (Figure 4 and 5) and Only patterns (Table 4 and 5) for agent-goal/object-subject relationship were then generated to examine the contextual role representations of the two words in Taiwan and Mainland Mandarin.

The common patterns (Figure 4 and 5) in two different regions reveal that the two node words seem to share more semantic common ground in their Mainland usage as both of their collocated subjects and objects in XIN have more commonality than those in the CNA corpus. In XIN, there are many collocations with political implications (e.g., 政党, *zheng4dang3*, political parties; 领导人 *ling3dao3 ren2*, political leaders) found in common patterns; while in CNA, only words bearing no social or political implication (i.e., 我方, *wo3fang1*, our party; 双方, *shuang1fang1*, two parties) were found to share common collocational patterns between the two words. It implies that the two synonyms are interchangeably used in political contexts in Mainland usage.

Additionally, however, in Taiwan Mandarin, the two words are treated with different contextual implications as *xie2shang1* are almost exclusively used in political contexts (e.g. 政党, *zheng4dang3*, political parties; 朝野, *chao2ye3*, government), while *tan2pan4* can be used in a broader range of activities, including political, economic or social settings (e.g. 政治性, *zheng4zhi4xing4*, political attributes; 资方, *zi1fang1*, investor; 财产权, *cai2chan3quan2*, property rights). Such difference might be the result of several rounds of language education reforms in mainland China in the past few decades, leading to a more hybrid usage of near-synonyms in Mainland Mandarin (Mills, 1956; Sheridan, 1981).



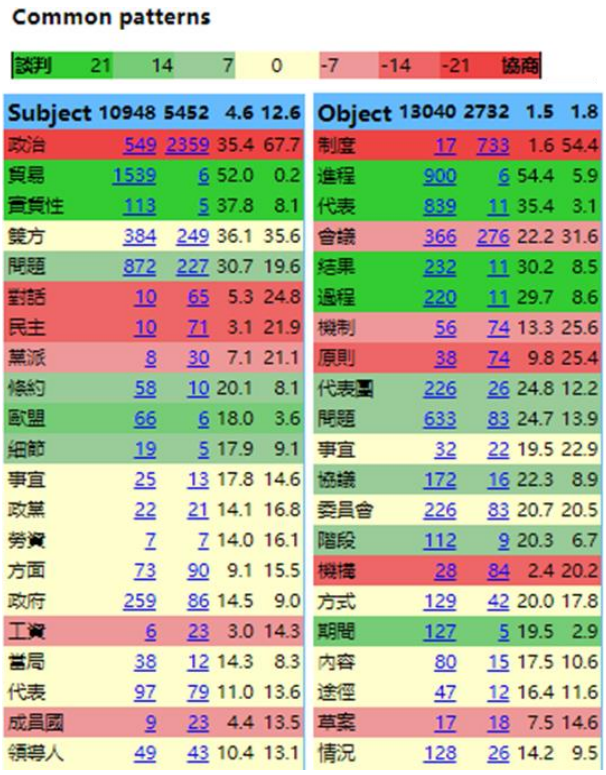
Common patterns

談判	21	14	7	0	-7	-14	-21	協商

Subject	13299	17548	4.4	11.8
朝野	6	4175	2.1	84.6
回合	804	7	64.4	4.9
事務性	23	318	19.9	55.6
政黨	16	910	4.9	48.9
政治性	146	28	45.3	21.2
兩黨	44	317	20.0	44.4
雙方	562	494	43.7	39.2
黨團	11	766	1.7	42.7
政治	870	342	38.8	23.2
在野黨	12	199	8.7	37.3
勞資	42	144	22.3	36.3
多黨	42	14	35.2	19.9
資方	48	25	29.8	20.6
財產權	105	5	29.4	2.7
部會	7	131	4.2	28.5
秘密	68	21	27.4	13.5
黨	88	244	17.4	27.2
美方	74	10	26.5	6.8
中共	498	144	26.3	10.2
我方	56	59	26.1	25.1
三通	53	14	25.2	10.8
馬拉松	36	10	25.1	11.3

Object	29986	11525	2.3	1.
大門	106	583	28.1	62.
籌碼	540	16	58.7	15.
結論	32	599	11.8	57.
管道	37	588	10.9	53.
機制	59	591	14.8	53.
代表	3402	135	48.1	13.
過程	907	265	42.7	32.
進展	382	12	41.5	8.
進程	235	12	41.3	11.
共識	31	351	6.2	39.
結果	835	88	38.3	17.
事宜	201	236	27.1	36.
底線	101	11	36.7	15.
僵局	177	12	35.1	11.
議題	356	247	32.1	34.
會議	215	815	10.1	34.
破局	14	23	21.5	31.
技巧	113	7	31.6	8.
策略	310	9	31.3	4.
空間	158	185	21.2	30.
協定	326	15	30.1	5.
協議	367	38	29.8	11.

Figure 4 Common patterns of *xie2shang1* and *tan2pan4* in CNA



Common patterns

談判	21	14	7	0	-7	-14	-21	協商

Subject	10948	5452	4.6	12.6
政治	549	2359	35.4	67.7
貿易	1539	6	52.0	0.2
實質性	113	5	37.8	8.1
雙方	384	249	36.1	35.6
問題	872	227	30.7	19.6
對話	10	65	5.3	24.8
民主	10	71	3.1	21.9
黨派	8	30	7.1	21.1
條約	58	10	20.1	8.1
歐盟	66	6	18.0	3.6
細節	19	5	17.9	9.1
事宜	25	13	17.8	14.6
政黨	22	21	14.1	16.8
勞資	7	7	14.0	16.1
方面	73	90	9.1	15.5
政府	259	86	14.5	9.0
工資	6	23	3.0	14.3
當局	38	12	14.3	8.3
代表	97	79	11.0	13.6
成員國	9	23	4.4	13.5
領導人	49	43	10.4	13.1

Object	13040	2732	1.5	1.8
制度	17	733	1.6	54.4
進程	900	6	54.4	5.9
代表	839	11	35.4	3.1
會議	366	276	22.2	31.6
結果	232	11	30.2	8.5
過程	220	11	29.7	8.6
機制	56	74	13.3	25.6
原則	38	74	9.8	25.4
代表團	226	26	24.8	12.2
問題	633	83	24.7	13.9
事宜	32	22	19.5	22.9
協議	172	16	22.3	8.9
委員會	226	83	20.7	20.5
階段	112	9	20.3	6.7
機構	28	84	2.4	20.2
方式	129	42	20.0	17.8
期間	127	5	19.5	2.9
內容	80	15	17.5	10.6
途徑	47	12	16.4	11.6
草案	17	18	7.5	14.6
情況	128	26	14.2	9.5

Figure 5 Common patterns of *xie2shang1* and *tan2pan4* in XIN

A further scrutinization of the object-subject relationship for only patterns of these two words across CNA and XIN echoed the above findings and revealed more clues for contextual usage (see Table 4 and 5).

In CNA, *tan2pan4* can be found in both economic and diplomatic contexts (e.g. 埃雷卡特 Saeb Erakat, a Pakistan diplomat). Additionally, further examination of their concordance lines confirms such findings, such as 新回合谈判 (*xin1hui2he2 tan2pan4*, new round of negotiation). However, *xie2shang1* is almost exclusively related to political contexts (see Table 4).

Table 4 Only patterns of *tan2pan4* and *xie2shang1* in CNA

tan2pan4			
Subject	**Freq.**	**NF.**	**MI**
贸易	1383	123.90	47.4
首席	439	39.33	53
航权	306	27.41	57.1
新回合	201	18.01	60.6
世界贸易组织	115	10.30	29.6
Object	**Freq.**	**NF.**	**MI**
代表团	500	44.80	37
龙永图*	200	17.92	53.8
对手	167	14.96	29.2
埃雷卡特#	132	11.83	55
程序性	88	7.88	42.1
	Total freq.		=111,619
xie2shang1			
Subject	**Freq.**	**NF.**	**MI**
人民	582	63.26	32.7
单位	372	40.44	23.8
党政	265	28.80	42.1
三党	83	9.02	32.7
预备性	23	2.50	26.2
Object	**Freq.**	**NF.**	**MI**
制度	120	13.04	19.7
版本	102	11.09	34.8
总预算案	85	9.24	32
会报	61	6.63	23.5
修正案	35	3.80	16.9
	Total freq.		=91,998

*(*a Chinese economist; # a diplomatic representative for Pakistan)*

In XIN, *tan2pan4* is primarily found in diplomatic contexts with less apparent reference to economic activities, while *xie2sahng1* can be found in both political and legal contexts (see Table 5). Additionally, objects for *xie2shang1* in Mainland usage can emphasize the exchanges of discussion being conducted with a possible mediator during such a process (e.g. 座谈会, *zuo4tan2hui4*, discussion panel; 委员, *wei3yuan2*, committee member).

Table 5 Only patterns of *tan2pan4* and *xie2shang1* in XIN

tan2pan4			
Subject	**Freq.**	**NF.**	**MI**
回合	475	70.58	58.8
首席	454	67.46	56.3
地位	453	67.31	40.4
入盟	320	47.55	61.5
阶段	227	33.73	30.3
Object	**Freq.**	**NF.**	**MI**
进展	286	42.50	34.1
立场	206	30.61	31
埃雷卡特#	223	33.13	59.6
龙永图*	133	19.76	52.1
僵局	110	16.34	37.2
	Total freq.		=67,301
xie2shang1			
Subject	**Freq.**	**NF.**	**MI**
人民	197	97.45	18.3
部门	134	66.29	18.7
单位	39	19.29	11
当事人	37	18.30	27.7
事务性	17	8.41	25.4
Object	**Freq.**	**NF.**	**MI**
委员	49	24.24	19.6
对话	23	11.38	17.2
职能	22	10.88	18
座谈	19	9.40	21.7
座谈会	19	9.40	21.7
	Total freq.		=20,215

Being loan words, such diverse contextual usage settings of *xie2shang1* and *tan2pan4* used in Mainland and Taiwan Mandarin might point to a sociological *status quo*, requiring further studies in disciplines of historical linguistics. It is also worth noting that no specific markers bearing role-internal attributes were found in their collocation across XIN and CNA.

4 Conclusion

Xie2shang1 and *tan2pan4*, as loan words, share synonymous common grounds for “reaching to some conclusions through discussions” with parties involved in such a process. Nonetheless, a comparative study of the two near-synonyms based on CNA and XIN of the CWS corpus reveals some distributional, eventual, and contextual similarities as well as differences across Taiwan and Mainland Mandarin.

Their distributional patterns suggested that, compared with the significantly underused *tan2pan4*, *xie2shang1* is significantly overused in Taiwan Mandarin, and vice versa in Mainland Mandarin. On their event representations, distinctions were not found between Mainland and Taiwan Mandarin as both words can refer to an inchoative process, though the starting point for *xie2shang1* is rather blurring compared with that in Taiwan Mandarin. It might be in relation to their relatively ‘young’ status within modern Chinese vocabulary.

Nonetheless, an examination of the subject-object/ agent-goal relationship of the two words revealed more details on their different contextual usages in Mainland and Taiwan Mandarin. The two words share more semantic common ground, or, more precisely, in political contexts of Mainland usage, *tai2pan4* and *xie2shang1* are used in a hybrid manner. Additionally, in Mainland Mandarin, *tan2pan4* is found more prominently used in diplomatic contexts with less apparent reference to economic activities, while *xie2sahng1* can be found in both political and legal contexts, emphasizing a possible mediator. In contrast, the object-subject relationship in Taiwan Mandarin suggests the two words are used in quite a different context. In Taiwan Mandarin, *tan2pan4* can be used in economic and diplomatic contexts, while *xie2shang1* is used exclusively in political contexts.

The role of contextual and distributional differences might point to other historical or sociological factors that may play a certain role in decerning the variational linguistic changes throughout time and history. Nonetheless, this does not fall into the role of the present study. Further investigations on a diachronic basis might contribute to our understanding of how and what loan words might evolve or change throughout historical, social, or political events.

Acknowledgements

The work has been funded by the 2019 Research Project for Young Scholars of Guangzhou Xinhua University (2019KYQN11; On the Collaborative Effects of Working Memory and Background Information on L2 Reading), and the Higher Educational Reform Project of Guangzhou Xinhua University (2019T001; Construction of Translation Teaching Team in School of Foreign Languages)

References

Chen, H. (2014). *A study of Japanese loanwords in Chinese* (Master's thesis).

Chief, L.-C., Huang, C.-R., Chen, K.-J., Tsai, M.-C., & Chang, L.-L. (2000). What can near synonyms tell us? *International Journal of Computational Linguistics & Chinese Language Processing*, *5*, 47-60. https://doi.org/10.30019/IJCLCLP.200002.0003

Chung, S. F., & Ahrens, K. (2008). MARVS revisited: Incorporating sense distribution and mutual information into near-synonym analyses. *Language and Linguistics: Lexicon, Grammar and Natural Language Processing*, *9*(2), 415-434.

Gunn, E. (1991). *Rewriting Chinese: Style and innovation in twentieth-century Chinese prose*. Stanford university press.

Hong, J. F., & Huang, C. (2008). A corpus-based approach to the discovery of cross-strait lexical contrasts. *Language and Linguistics*, 9(2), 221-238.

Huang, C. R., Ahrens, K., Chang, L. L., Chen, K. J., Liu, M. C., & Tsai, M. C. (2000, February). The module-attribute representation of verbal semantics: From semantic to argument structure. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 5, Number 1, February 2000: Special Issue on Chinese Verbal Semantics* (pp. 19-46).

Huang, C.-R., & Hsieh, S.-K. (2015). Chinese lexical semantics. In *The Oxford handbook of Chinese linguistics* (pp. 290-305). Oxford University Press.

Huang CR., Wang X. (2020) From Faithfulness to Information Quality: On 信 in Translation Studies. In: Lim L., Li D. (eds) *Key Issues in Translation Studies in China. New Frontiers in Translation Studies*. Springer, Singapore. https://doi.org/10.1007/978-981-15-5865-8_6

Hong, J. F., Huang, C. R., & Xu, M. W. (2007). 以中文十億詞語料庫為基礎之兩岸詞彙對比研究 (A Study of Lexical Differences between China and Taiwan based on the Chinese Gigaword Corpus)[In Chinese]. In *ROCLING 2007 Poster Papers* (pp. 287-301).

Hong, J. F., & Huang, C. (2006). Using Chinese gigaword corpus and Chinese word sketch in linguistic research. In *PACLIC 20 - Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation* (pp. 183-190)

Ahrens, K., Huang, C-R., & Chuang, Y. (2003). Sense and Meaning Facets in Verbal Semantics: A MARVS Perspective. *Language and Linguistics*, *4*(3), 468-484.

Kim, T. E. (2018). Mandarin loanwords. Routledge. https://doi.org/10.4324/9781351253406

Liang, Q., & Huang, C. R. (2021). Spatial-temporal attributes in verbal semantics: A corpus-based lexical semantic study of discriminating Mandarin near-synonyms of "tui1" and "la1". In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation* (pp. 314-321).

Ma, W. Y., & Huang, C. R. (2006, May). Uniform and effective tagging of a heterogeneous giga-word corpus. In *5th International Conference on Language Resources and Evaluation (LREC2006)* (pp. 24-28).

Masini, F. (1993). The formation of modern Chinese lexicon and its evolution toward a national language: the period from 1840 to 1898. *Journal of Chinese Linguistics monograph series*, (6), i-295.

Ma, W. Y., & Huang, C. R. (2006, May). Uniform and effective tagging of a heterogeneous giga-word corpus. *In 5th International Conference on Language Resources and Evaluation (LREC2006)* (pp. 24-28).

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Mills, H. (1956). Language Reform in China: Some Recent Developments. *The Far Eastern Quarterly*, 15(4), 517-540. https://doi.org/10.2307/2941922

Sheridan, E. M. (1981). Literacy and Language Reform in the People's Republic of China. *The Reading Teacher*, *34*(7), 804–808. http://www.jstor.org/stable/20195341

Shi, Y. (2020). Loanwords in the Chinese Language. Routledge. https://doi.org/10.4324/9781003131359

Shi, Y.-W. (2019). Xinhua Loanwords Dictionary (新华外来词词典). The Commercial Press.

Tao, Y. (2021). Study of Near Synonymous Mental-State Verbs: A MARVS Perspective. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation* (pp. 436-444).

Wang, H., & Zhang, Y. (2010). Deliberation and Consultation: an important operational form of demographic politics (协商合作：一种重要的民主政治运作形式）. *Academic Journal of the China Institute of Theory on the Chinese People's Political Consultative Conference*, (02), 33-39.

Wang, X., & Huang, C.-R. (2018). From near-synonyms to power relation variations in communication: A cross-strait comparison of "guli" and "mianli". In *Workshop on Chinese Lexical Semantics* (pp. 155-166). Springer.

Wang, Z., & Huang, C. R. (2021). From Near-synonyms to Divergent Viewpoint Foci: A Corpus-based MARVS driven account of two verbs of attention. In *Proceedings of the 35th Pacific Asia Conference* *on Language, Information and Computation* (pp. 282-290).

Zhan, W., Guo, R., Chang, B., Chen, Y., & Chen, L. (2019). The building of the CCL corpus: Its design and implementation. *Corpus Linguistics*. 6 (1), 71-86.

How syntactic analysis influences the calculation of mean dependency distance: Evidence from the enhanced dependency representation

Tsy Yih
Department of Linguistics
Zhejiang University
yezi_leafy@hotmail.com

Jianwei Yan
Department of Linguistics
Zhejiang University
yanjianwei@aliyun.com

Haitao Liu✉
Department of Linguistics
Zhejiang University
lhtzju@gmail.com

Abstract

Based on Yan & Liu (2022), the present paper further explores how syntactic analysis, or the annotation scheme of dependency treebanks, affects mean dependency distance (MDD). By comparing the treebanks of 16 languages with both basic (BUD) and enhanced universal dependency (EUD) representations, we find that the MDD measured by the EUD representation is statistically larger than that of BUD. The main distinction between the two representations lies in the treatment of three constructions: coordinate structures, relative clauses, and pivotal constructions. However, a closer look at the data reveals that these three analyses in EUD do not necessarily have longer MDD in single sentences: The enhanced analysis of coordinate structures and pivotal constructions statistically have a significant contribution to the increase of MDD, while that of relative clauses contributes the least. We conclude with the factors that may affect the change of MDD, including both internal, structural ones and external ones, such as the context of the text (in the form of stochastically intervening dependents) and the language type (in the form of word order type).

1 Introduction

Mean Dependency Distance (MDD), defined as the sum of all dependency distances divided by the number of dependency relations, is a measure based on the dependency structures of sentences. It has received much attention in the last two decades (Liu, 2008; Futrell et al., 2015; Jiang & Liu, 2015). Previous studies have shown a general tendency for natural languages to possess statistically smaller MDD than languages generated by a number of random baselines (Liu, 2008; Futrell et al., 2020). Therefore, it is generally considered to be a metric to reflect syntactic complexity and the limit of human memory, and many studies have attempted to explain its inner mechanism (Temperley, 2008; Gildea & Temperley, 2010; Liu et al., 2017), such as being the result of the Principle of Least Efforts (Zipf, 1949). MDD is known to be subject to many factors, such as language type (Liu, 2008), sentence length (Jiang & Liu, 2015), chunking (Lu et al., 2016), genre (Wang & Liu, 2017), annotation scheme (Yan & Liu, 2022), etc.

Among all these factors, the annotation scheme differs from others in that it influences the observed value of the MDD, rather than the real value of the variable itself. An analogy is to measure the temperature with different scales, and one would have different values. For instance, a certain temperature might be measured to have the value 40 under degree Celsius, and show the value of 104 under degree Fahrenheit. Turning back to the linguistic issue here, an annotation scheme reflects the choice of syntactic analysis. Since there are various versions of syntactic structural analysis in the linguistic literature, it is thus also important to pay attention to such effect. As we all know now, an underlying formula exists for the abovementioned case of temperature scales: $(C \times 9/5) + 32 = F$. Likewise, in studying MDDs under different annotation schemes, one aim is also to find such relationship, although as we will show below, it is not that easy to have a function relation for the linguistic case.

Previously, Yan & Liu (2022) have made systematic investigations into how the syntactic annotation scheme affects the calculation of dependency distance by comparing UD and Surface-Syntactic Universal Dependencies (SUD), and they found that the MDDs in SUD are statistically shorter than those in UD. In their study, the four major constructions or controversial pairs where the dependency structures are different in two annotation schemes are the adposition-noun, auxiliary-verb, copula-noun/adjective, subordinator-verb pairs. In general, UD takes a content-head approach and SUD a function-head approach. For instance, in UD an adposition is the dependent of the head noun, while in SUD it is the head of the noun and serves as the linker between the verb and the noun. Hence, if the human language generally follows the Principle of Relator Being Intermediate (Dik, 1997), then the SUD analysis is bound to have shorter MDD values. From the comparison between UD and SUD have we learned that an annotation scheme can be seen as the combination of analyses of various linguistic constructions, and that the analysis of different constructions might probably have conflicting effects on MDD, making it much more complex than the one-dimensional variable of temperature. Yet, at least it would be worth accumulating more case studies in this trend at this stage.

Among all variants of annotation schemes available now, the enhanced dependencies are noteworthy. The enhanced representation was couched in de Marneffe et al. (2014) since the time of Stanford Dependencies (SD), and was later succeeded by the Universal Dependencies (UD) Initiative (Nivre et al., 2016). [1] In contrast with the basic dependencies which only allow tree structures, the enhanced representation allows graph structures or cyclic parts, and supplements additional relations.[2] Schuster and Manning (2016) later proposed a version of enhanced and enhanced++ UD representations,[3] which will be together called EUD in the present study, in contrast with the basic universal dependencies (BUD) representation.[4]

Taking the sample sentence in Table 1 as an instance, the 9th column (DEPS) of line 5 (dogs) in the enhanced format would be "2:obj|3:conj:and" rather than "3:conj", which indicates two relations. The graphic syntactic structures of the sentence in two formats are shown in Figure 1.

It would be interesting if we extend the calculation of MDD from BUD to EUD. With the additional links in the enhanced representation, the mean dependency distance of sentences and the whole treebanks, by definition, is subject to change. However, since both the number of relations and the sum of all dependency distances have changed, it is unclear whether MDDs would increase or decrease.

Hence, we put forward the following research questions:

(1) Do the enhanced MDDs increase or decrease compared with the original MDDs?

(2) What factors lead to the change of MDD in enhanced representation?

[1] The term "UD" could be ambiguous. In one sense, it refers to a specific annotation scheme, i.e., Yan & Liu's UD contrasted with SUD, or the BUD contrasted with EUD in the present study. In another sense, it stands for the whole annotation initiative (Zeman et al., 2017) following a specific format *.conllu*, which already has 202 treebanks of 114 languages till v2.8 (https://universaldependencies.org/).

[2] We simply use the term "relations" or "links" rather than "dependency relations" as it is hard to say if there is superiority between two words.

[3] For more information, the reader can also refer to https://universaldependencies.org/u/overview/enhanced-syntax.html.

[4] Punctuations are generally not included in the calculation of MDD.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
# text = I love cats and dogs.									
1	I	I	NOUN	PRP	_	2	nsubj	2:nsubj	_
2	love	love	VERB	VBP	_	0	root	0:root	_
3	cats	cat	NOUN	NNS	_	2	obj	2:obj	_
4	and	and	CCONJ	CC	_	5	cc	5:cc	_
5	dogs	dog	NOUN	NNS	_	3	conj	3:conj	_
6	.	.	PUNC	.	_	2	punct	2:punct	_

Table 1. The conllu format of a sample sentence

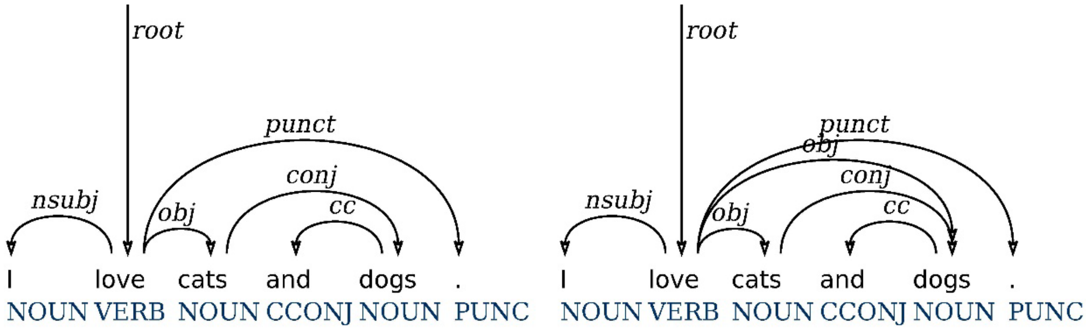


Figure 1. The basic UD annotation (the left panel) and its enhanced version (the right panel)

Language	Genera	Corpus	Genre	Sentences	Tokens$_{sp}$[5]
Arabic	Semitic	PADT.test	News	672	25911
Belarusian	IE, Slavic	HSE.test	Mixed	1020	12935
Bulgarian	IE, Slavic	BTB.test	Fiction, legal, news	1111	13451
Czech	IE, Slavic	PUD	News, wiki	984	15737
Dutch	IE, Germanic	LassySmall.test	Wiki	765	9441
English	IE, Germanic	PUD	News, wiki	993	18609
Estonian	Uralic, Finnic	EWT.test	Blog, social, web	866	10404
Finnish	Uralic, Finnic	TDT.test	Mixed	1515	17463
Italian	IE, Romance	ISDT.test	Legal, news, wiki	481	9217
Latvian	IE, Baltic	LVTB.test	Mixed	1852	23108
Lithuanian	IE, Baltic	ALKSNIS.test	Mixed	671	8774
Polish	IE, Slavic	PUD	News, wiki	1000	15731
Slovak	IE, Slavic	SNK.test	(Non-)Fiction, news	1013	10677
Swedish	IE, Germanic	PUD	News, wiki	993	17025
Tamil	Dravidian	TTB	News	600	8581
Ukrainian	IE, Slavic	IU.test	Mixed	816	13227

* The treebanks with more than three subgenres were recorded as having a mixed genre in the table.

Table 2. Basic information of the treebanks

[5] The subscript $_{sp}$ is the shorthand for *sans punctuation*, i.e., without punctuations, contrasted with $_{cp}$ – con punctuation. We devise these abbreviations for the author to report clearly and for the reader to obtain the correct information about the corpora directly. The Latin prepositions are preferred over *with* and *without* in English since they have the same initial letter "*w*".

2 Material and Methods

2.1 Material

We picked out all of the language samples that have enhanced dependencies and, most importantly, multiple dependencies from the Universal Dependencies initiative [6] . [7] As a result, 16 languages remained, as shown in Table 2. We selected the test sets of these languages to have a controllable size.[8] The only exception is Tamil's TTB: We included all of its test, training, and development sets to ensure that the size of the treebank is comparable with other treebanks.

2.2 Methods

For calculating the dependency distance of an enhanced dependency treebank, we shall start by calculating the dependency distance of a sentence. Here, Liu's (2008) approach was adopted. Formally, let $w_1 ... w_i ... w_n$ be a word string of length $_n$. For any dependency relation between the words w_x and w_y ($x \geq 1, y \leq n$), if w_x is a head and w_y is its dependent, then the dependency distance (DD) between them is defined as the absolute value of the difference $|x - y|$. Therefore, the mean dependency distance (MDD) of a sentence is defined as:

$$MDD(\text{sentence}) = \frac{1}{n-1}\sum_{i=1}^{n-1}|DD_i| \qquad (1)$$

where n is the number of words in a sentence and DD_i is the dependency distance of the i-th dependency relation of the sentence. Another way to define the sentential MDD is as follows:

$$MDD(\text{sentence}) = \frac{1}{m}\sum_{i=1}^{m}|DD_i| \qquad (2)$$

where m is the number of all dependency relations, which is probably either equal to or larger than n−1. Note that there is a separate line for the root node in *conllu* format, whereas its dependency distance is zero.

Based on the second definition, the MDD of the whole treebank can be defined as:

$$MDD(\text{treebank}) = \frac{1}{M}\sum_{i=1}^{M}|DD_i| \qquad (3)$$

where M is the whole relations in a treebank, which is equal to the sum of relations in each sentence. By such definition, the MDDs in both basic and enhanced representations are each a special case. In doing so, it is compatible and comparable for both cases. According this formula, the MDD of the example sentence *I love cats and dogs* in the last section in BUD representation is (1 + 1 + 2 + 1) / 4 = 1.25, while the MDD of the same sentence in EUD representation is (1 + 1 + 2 + 1 + 3) / 5 = 1.6. In this case, the MDD of EUD is higher than that of BUD.

We processed the treebank in Microsoft Excel by importing the *.conllu* format treebanks into worksheets, and did the statistical analysis by the R language[9]. The procedure of our data processing is as follows: We first deleted three kinds of sentences: The first kinds includes those which only have one root word except for punctuations, where the dependency distance cannot be calculated. The second kind contains sentences where punctuations are heads of other tokens, which causes problems when deleting punctuations. These sentences were deleted because they are hard to deal with if the language is unintelligible to us. The third kind consists of those with empty nodes because there is divergence in the treatment of the position of the empty node. For instance, the English PUD treebank duplicates the node right after the original node, while in other treebanks, the empty node can appear in any supposed position in the sentence. Yet the calculation of dependency distance relies on the exact position of words. Hence the indeterminacy of the surface position of empty nodes could be problematic. The second step was to convert all the values in the ID and HEAD columns into a relative reference in Excel for the ease of the next step. The third step was then to delete all the punctuations, a treatment

[6] All the treebanks are available from https://universaldependencies.org/.

[7] The descriptions of some treebanks claim to have enhanced dependencies, while they are simply the copy of basic dependencies without additional links.

[8] All the PUD (Parallel Universal Dependencies) treebanks only have the *test* set. Therefore they were not elucidated in the table.

[9] https://www.r-project.org/.

following Jiang & Liu (2015) and other previous studies for comparison. Since the position numbers are now relative references, they will change automatically after the rows containing punctuations were deleted. Then we calculated MDDs for both the basic and enhanced representations. Finally, the results were exported and put into R for statistical analysis if necessary.

3 Results and Discussion

3.1 Enhanced MDDs Compared with Basic MDDs

Table 3 shows the MDDs of 16 languages in both BUD and EUD annotation schemes. It can be seen that in all languages the enhanced MDDs are higher than basic MDDs, although the increments in each language are different. A paired one-sided Wilcoxon test shows that the enhanced MDD is significantly greater than the basic MDDs ($V = 0$, $p = 1.526\text{e-}05 < 0.05$).

Theoretically, the enhanced MDD is not bound to be larger than the basic MDD. When we add a new link, if the new link is an adjacent one or it has a smaller dependency distance than the original MDD of that sentence, then the whole MDD is supposed to decrease by maths. This is not a rare thing since Liu (2008), Jiang & Liu (2015) have found that adjacent relations are very common in natural language and would take up about 50% of the whole dependency relations. Futurell (2019) also argued that adjacent relation, or so-called information locality, is preferred in the structuring of language. If a new link is added by chance, then it is very likely to be adjacent. Our results, however, have revealed that the MDD goes up, indicating that the additional relations are generally long-distance ones rather than adjacent ones. To put it differently, the results seem to show that the original annotation scheme itself tends to adopt an analysis that keeps the short dependencies and omits the long-distance ones.

However, it is noteworthy that it is just statistically the EUD representations manifest longer MDDs, while there are also many single sentences with shorter MDDs, such as in (1) where the reanalysis of relative clauses in EUD plays a part.

(1) For those who follow social media transitions on Capitol Hill, this will be a little different. (English PUD)

BMDD: 2.8667

EMDD: 2.8125

A second aspect worth mentioning is that if we arrange the table in an ascending or descending order according to the MDDs before and after enhancement, the languages will be in different orders, which coincides with Yan & Liu (2022)'s finding in comparing the UD and SUD annotation scheme. This indicates that the MDD is affected by both annotation scheme and language type (e.g. head-final or head-initial). Otherwise, the orders in different representations should be the same. Hence, it is the interaction of these two factors that decide the value of MDDs. What is the nature of annotation scheme then and what part does it play in determining MDDs?

In the next section, we take a closer look at the distinction between two representations and explore what constructions have led to the increase of MDDs.

3.2 The Constructions Contributing to the Change of MDDs

As the results in the last section have indicated that the enhanced MDDs are longer than the basic MDDs in most languages, it is then natural to inquire what factors have contributed to the increase of MDDs.

Similar to Yan & Liu (2022), we decomposed two annotations schemes into constructions of which they have different analyses. The four previous phenomena do not have a distinct analysis in EUD and the cycles are not recovered in the enhanced analysis. We had a different set of constructions. Table 4 shows all types of enhancement of EUD given by Schuster and Manning (2016).

Language	Basic MDD	Enhanced MDD	Increase

Arabic	3.195	3.806	19.12%
Belarusian	2.414	2.758	14.25%
Bulgarian	2.304	2.455	6.55%
Czech	2.391	2.508	4.89%
Dutch	2.676	2.895	8.18%
English	2.528	2.715	7.40%
Estonian	2.644	2.647	0.11%
Finnish	2.307	2.633	14.13%
Italian	2.519	2.787	10.64%
Latvian	2.442	2.806	14.91%
Lithuanian	2.492	2.778	11.48%
Polish	2.226	2.415	8.49%
Slovak	2.102	2.246	6.85%
Swedish	2.473	2.647	7.04%
Tamil	2.399	2.428	1.21%
Ukrainian	2.625	3.011	14.70%

Table 3. The basic MDDs and enhanced MDDs in 16 languages

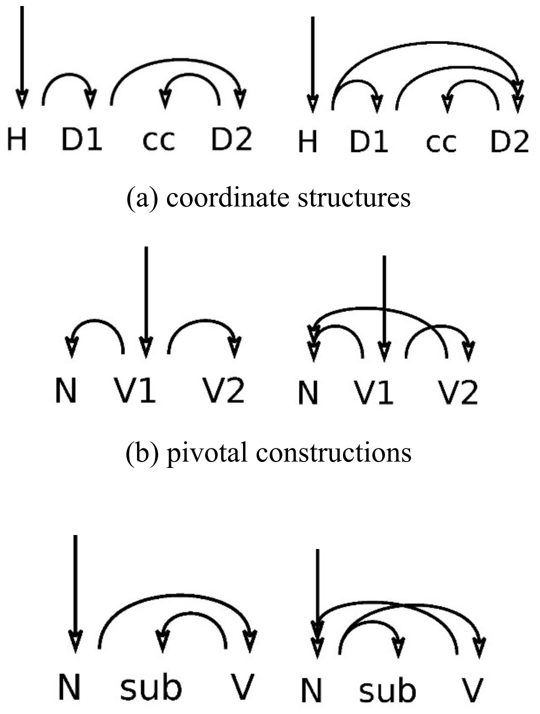
Version	Types of enhanced dependencies	Affecting MDD?
The enhanced UD representation	Augmented modifiers	No
	Augmented conjuncts	No
	Propagated governors and dependents	Yes
	Subjects of controlled verbs	Yes
The enhanced++ UD representation	Partitives and light noun constructions	Yes
	Multi-word prepositions	No
	Conjoined prepositions and prepositional phrases	Yes
	Relative pronouns	Yes

Table 4. Types of enhanced dependencies and their effects

Several types of additional relations above only elaborate on the relations but do not change the dependency distances, such as those with the value of "No" in the last column. As for the rest of them, the so-called "partitives and light noun constructions" in their treatment are noted as having the dependency relation *qmod*. Yet we have not found the *qmod* relation in any annotated treebank. Besides that, the case of "conjoined prepositions and prepositional phrases" concerns empty nodes. As we have deleted the sentences with empty nodes in processing the data since they are not treated equally in different languages and might cause problems, they will not be of our concern in the present study. Therefore, the three remaining primary types of enhanced relations left are associated with coordinate structures, pivotal constructions [10] and relative clauses, which correspond to "propagated governors and dependents", "subjects of controlled verbs" and "relative pronouns", respectively, in the original terms.

Put another way, the EUD and BUD are decomposed into the combination of different analyses of these three constructions. In what follows, we will first demonstrate their treatment in two representations and then see how they affect MDDs.

[10] The more commonly used term in the Western literature is controlled and raising structure, while we follow the use of pivotal constructions as in the Chinese linguistic literature here (Peng, 2017). In this case, the upper-level verb takes another verb as one of its syntactic argument, thereby rendering one semantic argument of the lower-level verb disappear. One has to trace its referent from the arguments of the upper-level verb. Prototypical cases include *want to*, *need to*, *start* and so forth.



(a) coordinate structures

(b) pivotal constructions

(c) relative clauses

Figure 2. The analyses of coordinate structures, pivotal constructions and relative clauses in BUD (left panel) and EUD (right panel)

As can be seen in Figure 2, in terms of the first two constructions, the enhanced structures are supergraphs of the basic structures. That is, the EUD analysis only has additional links, reflecting referential relations or functional equivalence. Hence, for the mathematical rationales presented in Section 3.1, it is easy to predict whether MDD increases or decreases based on the length of the additional links. If the length of the additional relation is longer than the original MDD in the basic setting, then it will increase the MDD in the EUD representation. However, in the case of Figure 2 (c), the relative clauses are more complicated as there are not only additional links but also changes of old links. On the one hand, there is an additional relationship between the antecedent and the root in the subordinate clause, forming a mutual dependency. On the other hand, the head of the relative pronoun is changed from the subordinate root to the antecedent. The change of DD in this local structure is |NV| + |Nsub| – |Vsub|. In the simplest case where these three elements form a continuous sequence, as |Nsub| = |Vsub| = 1, and |NV| = 2, the overall amount of increase of DD is 2. As can be seen from the graphs, the least increase of MDD in the three cases are 3, 2, and 2. Liu (2008) has shown that the MDDs in most languages fall between 2 and 3, which indicates that coordinate structures are very likely to contribute to the increase of MDDs, while the latter two constructions might probably decrease MDDs. Since the analysis above is purely theoretical and the MDD of a specific sentence may vary and is subject to the sentence length, we computed the proportions of how on earth these three constructions affect the MDD of all 16 languages dynamically, as shown in Figure 3.

In Figure 3, the black parts are those contributing to the increase of MDDs, while the white ones are those leading to the decrease of MDDs. A first sight suggests that all three constructions have the possibility to both increase and decrease MDDs, indicating that the competition between enlargement and reduction of MDD is dynamic rather than absolute.

Next, we hypothesized that the EUD analysis of the coordinate structures increases MDDs while the latter two decrease MDDs. However, the results indicate a general tendency for each construction to have an increased MDD. The overall situation does confirm that the coordinate structures have a large contribution, while relative clauses do possess less proportion. However, in many languages, the increasing part is still larger than the decreasing one (as shown by those black parts that take up more than 0.5 of all such constructions). As for the pivotal constructions, most of them increase the MDD. The one-sample sign test shows that the medians of the *conj* and *pivot* groups are greater than 0.5 (S =14, p = 0.0005 < 0.05), while that of *rel* is not significantly different from 0.5. The rank sum test shows there is no significant difference between *conj* and *pivot*, whereas both of them are greater than *rel* by one-sided tests (W = 172, p = 3.854e-06 for *conj* and *rel*, W = 167, p = 2.14e-05 for *pivot* and *rel*).

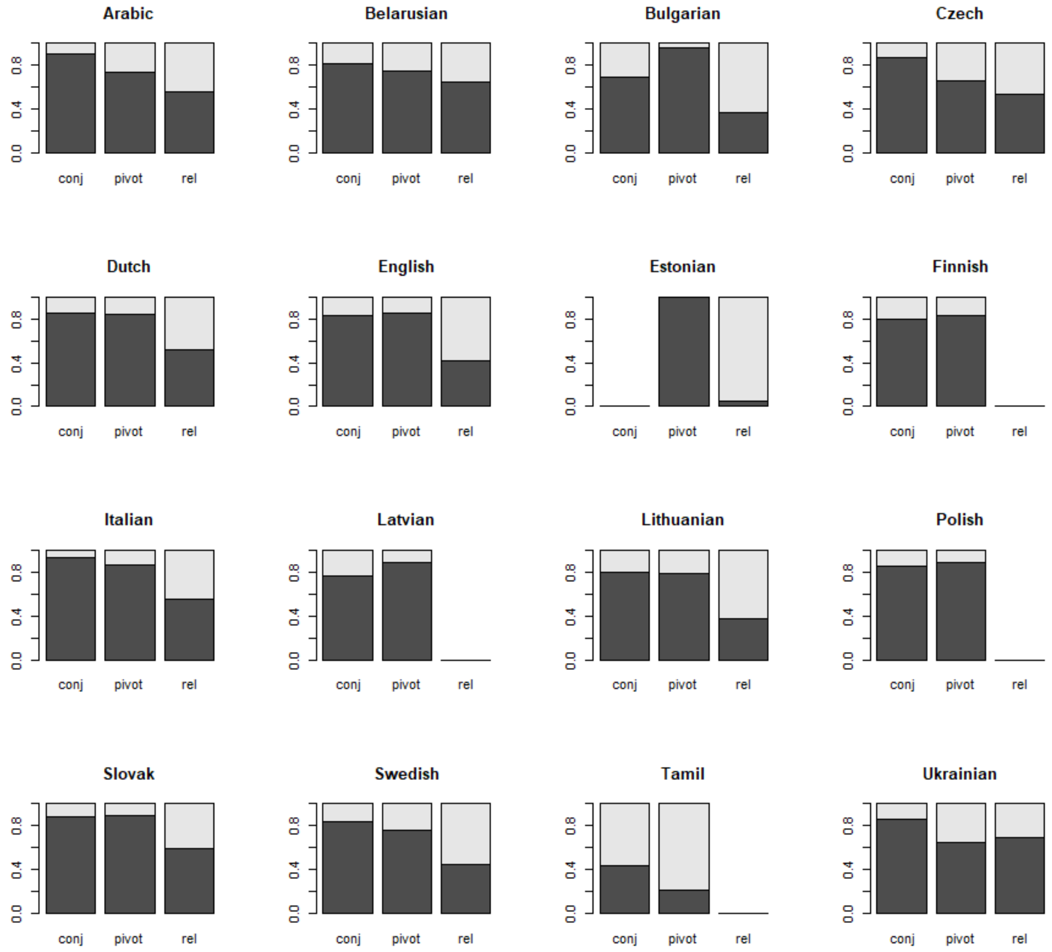


Figure 3. The proportions of the three categories in the 16 languages (black: increase; white: decrease)

Since our analyses above are based on the simplest and ideal cases, there must be other factors to be taken into consideration. The reasons can be from several aspects. From the internal, structural perspective, we have assumed a [N sub V] configuration for relative clauses where the antecedent noun, the subordinator (i.e. the relativizier or relative pronoun) and the root verb in the subordinate clause form a continuous sequence. Nevertheless, this only happens in a few restricted cases, where the verbs do not appear at the end of the subordinate clause, and the antecedents serve as the subject which is supposed to be at the beginning of the sentence. In other situations, for instance, the antecedents play the role of objects or obliques, then the words lying between them might rise dramatically. The same holds for pivotal constructions. Our analysis above is ideal and do not consider the cases such as *want to*. Even one additional particle such as *to* here would make the increase of MDD of the local structure at least to 3, which makes it probable to exceed the original MDD.

Other external factors include the content of the text and the language type. A closer look at the data reveals that there are many intervening tokens or dependents. Since the EUD representations are graph-based and have additional relations, if these links cross over a longer distance than the original MDD, they will give rise to its increase. These are not predicted from structural analysis but determined by the content that the addresser express.

Another possible factor is the language type. By language type we especially refer to the word order type. For instance, head-final languages are found to have longer dependency distances (Futrell et al., 2020: 397). In terms of the three constructions we concern here, in those languages where the subordinate clauses are verb-final, as the antecedent will be far from the subordinate root, the EUD treatment might probably lead to an increase. As for the UD analysis of coordinate structures, the head governs the first conjunct and then the latter the second conjunct, which is related to the linear sequence. However, in a head-final language, obviously such analysis would lead to longer MDD. One might also think of an alternative annotation scheme where head-final language has a shorter MDD, such as making the last conjunct connect to the head first. Overall, we can conclude that the interaction of annotation scheme and language type would affect the values of MDDs.

There are also some problematic data. It can be found that in some languages there is no such relation at all, which indicates an annotation difference. On the one hand, there is few *conj* relation in the Estonian treebank which is also problematic, as it is almost impossible to have no coordinate construction in a not-too-small corpus. On the other hand, in the Finnish, Latvian and Polish treebanks, the EUD annotation scheme does not deal properly with relative pronouns. However, there are indeed such words as *joka* (Finnish), *kas*, *kurš* (Latvian), and *który*, *jaki* (Polish). As for the case of Tamil, it employs an affix *-a* as the relativizer, which is not suitable for the relative pronoun analysis in those European languages. This suggests that the analysis of relative clauses in UD requires reconsideration. From a cross-linguistic perspective, many languages do need relativizers, but they might not be referential as English's so-called "relative pronouns" seem to be. Therefore relativizers might also be better treated as some subordinators as those in complement clauses[11] or as a separate category, as one of UD's goals is to maximize cross-linguistic parallelism or as Croft et al. (2017) have pointed out.

4 Conclusion

Thus far, the points to be made in the present study includes:

1. Empirically, the MDDs in the EUD representation are longer than those in the basic UD representation. Specifically, for all the three major distinctive constructions, there are cases where they increase or decrease MDD.
2. The EUD analysis of coordinate structures contributes most to the increase of MDDs, followed by that of pivotal constructions. Relative clauses, although on the whole also increase MDDs in the EUD representation, yet they have the strongest tendency to decrease among the three constructions.
3. The factors that lead to the changes include both internal, structural, and external ones, such as the content of the text (in the form of stochastically intervening dependents) and the language type (in terms of word order type). A more detailed investigation into the effects of these factors is beyond the scope of this paper and requires more comprehensive theoretical analyses and empirical validations.

To conclude, we want to re-emphasize the view that the nature of annotation scheme is the combination of analyses of various linguistic phenomena or constructions. Particularly, while the EUD representations seem to be redundant, it is simply one alternative analysis among the various possible dependency syntactic analyses. The present research is also supposed to deepen our understanding of the idea of "grammatical analysis as measurement" in language description.

A next step might be to compare more annotations schemes and decompose them into micro-parameters. Once we can manually calibrate each parameter at our will, we are likely to gain a deeper understanding of how the annotation scheme would affect the results of linguistic measurements.

Acknowledgments

This research was funded by the National Social Science Foundation in China (Grant No. 20CYY030) and the Humanities and Social Sciences Youth Foundation of the Ministry of Education of China (Grant No. 21YJC740060).

[11] This same goes to the marker of adverbials clauses such as *when* and *where*. In the current version of English UD, words like *before* and *after* are treated as subordinator but *when* and *where* are treated as adverbial modifiers, which are inconsistent.

References

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 26–31, Paris.

William Croft, Dawn Nordquist, Katherine Looney, Michael Regan. 2017. Linguistic typology meets universal dependencies. In *Proceedings of The Fifteenth International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75, Bloomington, IN.

Simon C. Dik. 1997. *The Theory of Functional Grammar (2nd ed.)*. Mouton de Gruyter, Berlin.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, 112:10336–10341.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of First Workshop on Quantitative Syntax*, pages. 1–15, Paris.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length?. *Cognitive Science*, 34(2): 286–310.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency Treebank. *Language Sciences*, 50:93–104.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Qian Lu, Chunshan Xu, and Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages?. *Complexity*, 21(S2):33–41.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldbergy, Jan Hajičz, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož.

Rui Peng. 2017. *Pivotal Constructions in Chinese: Diachronic, synchronic, and constructional perspectives*. Benjamins, Amsterdam.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož.

David Temperley. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59: 135–147.

Jianwei Yan and Haitao Liu. 2022. Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica*, 76(2): 406–428.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA.

Chat-log Disentanglement via Same-Thread Classification and Direct-Reply Prediction

Chia-Hui Chang, Zhi-Xian Liu
Yu-Ching Liao, **Yu-Hao Wu**
National Central University, Taiwan
chia@csie.ncu.edu.tw

Thamolwan Poopradubsil
Department of Computer Science
Kasetsart University, Thailand
tmw.poopradubsil@gmail.com

Abstract

The early purpose of chatlog (conversation) disentanglement is to separate intermingled messages into detached conversations for easier information following and relevant information retrieving from simultaneous messages. Thus, the problem has been modeled as predicting whether two messages come from the same-thread. While the previous study by (Jiang et al., 2018) seems to perform well on same-thread prediction, we find that it is because the data are randomly split into training and test sets, resulting overlapping of topics in training and testing sets. When data is split by time order, the performance of existing models drop significantly. In this study, we consider the problem of direct reply predication task and study different message pair classification models for the task. We argue that independent message encoders could better represent messages to capture their interaction than shared message encoders especially for direct-reply prediction task. We also find that BERT model performs well with small datasets, while other models may outperform BERT with large datasets.

1 Introduction

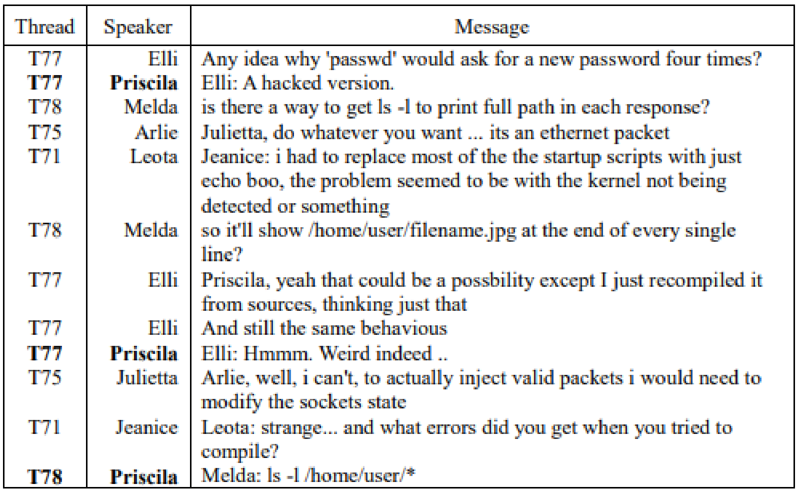
With the continuous development of the Internet and social media, online group discussions and conversations have become increasingly popular and play an important role in society and the economy. Many commercial sites take advantage of this to advertise their products or help solve users' problems. For example, developing conversational agents (chatbots) on their website to help staffs answer questions that may have been asked before.

The goal of chatlog disentanglement is to cluster messages that belong to the **same topic** for tracing. Selecting a reply for a given input or finding question-answer pairs are special cases of conversation disentanglement. For example, Figure 1 shows a segment of conversations consisting of four ongoing threads in the IRC (Internet Relay Chat) conversation. As we can see, interleaved conversations can occur in both two-person or multi-person chats. Thus, the goal of conversation disentanglement is to match the message pairs for question-answer pair generation.

There are many studies on conversation disentanglement in the past. One solution for conversation disentanglement is to model the topic of messages by estimating the similarity between messages and decide whether each incoming message starts a new topic or belong to an existing thread.

Jiang et al. (2018) proposed Siamese Hierarchical Convolutional Neural Network (SHCNN) which integrates two hierarchical CNNs to capturing low-order and high-order semantics of the messages for a better message representation. By concatenating absolute difference of the two representation with other temporal and user information, SHCNN predicts the probability of two messages belonging to the same-thread with high accuracy on the data of IRC and Reddit. However, the model performs poorly for future unseen messages when the training and testing data are split by time order.

We argue that the proposed data preparation method by (Jiang et al., 2018) only avoid the generation of too many negative examples, but may produce false positive message pairs as many subtopics fork from the main topic. When two messages from dif-



Thread	Speaker	Message
T77	Elli	Any idea why 'passwd' would ask for a new password four times?
T77	**Priscila**	Elli: A hacked version.
T78	Melda	is there a way to get ls -l to print full path in each response?
T75	Arlie	Julietta, do whatever you want ... its an ethernet packet
T71	Leota	Jeanice: i had to replace most of the the startup scripts with just echo boo, the problem seemed to be with the kernel not being detected or something
T78	Melda	so it'll show /home/user/filename.jpg at the end of every single line?
T77	Elli	Priscila, yeah that could be a possbility except I just recompiled it from sources, thinking just that
T77	Elli	And still the same behavious
T77	**Priscila**	Elli: Hmmm. Weird indeed ..
T75	Julietta	Arlie, well, i can't, to actually inject valid packets i would need to modify the sockets state
T71	Jeanice	Leota: strange... and what errors did you get when you tried to compile?
T78	**Priscila**	Melda: ls -l /home/user/*

Figure 1: Conversation in real-world chatting room

ferent subtopic are paired as positive examples, even humans have difficulty to recognize their relationship. In this paper, we consider a different task to pair only **direct-reply** messages for positive examples, synthesizing a more reasonable data set for question-answer pair extraction. We consider three neural networks models based on GloVe word embedding, including CNN+LSTM, LSTM with dual attention, and attention over attention (AOA) (Huang et al., 2018) and show improved performance over SHCNN. However, the best performance only achieves 0.669 F1, even with the BERT sentence pair classification.

In addition, we apply the direct-reply prediction task to extract question-answer pairs from chatlogs and find substantial labeling is required to obtain acceptable model. To speed up the process, we adopt heuristic labeling of the next sentence as a reply message to speed up training data preparation. Overall, chatlog disentanglement is still a challenging problem to be solved.

2 Related Work

The early research on conversation disentanglement can be traced back to the study on topic detection and tracking conducted in (Allan, 2002). As mentioned in (Shen et al., 2006), the messages in the same-thread have higher similarity. Thus, calculating message similarities based on linguistic features based on bag-of-words representation has been the major idea in (Elsner and Charniak, 2008). In addition, (Wang and Oard, 2009) showed that contexts can be used to improve the performance of message similarity calculations. However, overlapping contexts could also influence the calculation of similarity, leading to reduced performance.

Mehri and Carenini (2017) proposed a pipeline for the task of thread disentanglement, including reply classifier, same-thread classifier, next utterance classifier, and in-thread classifier. Of the three subtasks, only the third classifier, i.e. for “next utterance classification”, can leverage unlabeled data to model message relationships to train an Recurrent Neural Network (RNN) classifier (Lowe et al., 2015). Finally, the “in-thread” classifier takes the output of the previous three classifiers “same-thread”, “Reply”, and ”Next Utterance” to predict if an input message belongs to a thread.

To investigate how message similarity could be estimated, Jiang et al. (2018) proposed SHCNN (Siamese Hierarchical Convolutional Neural Networks) for the same-thread task prediction. They merged messages from different subReddits to simulate the concurrent conversations with multiple threads and generated a synthetic dataset of interleaved conversations, where messages from the same reddits within a limited elapsed time are paired to be positive examples, while messages from different reddits are paired to be negative examples. The experimental results show that the model performs well with 0.8392 MRR when the data are randomly

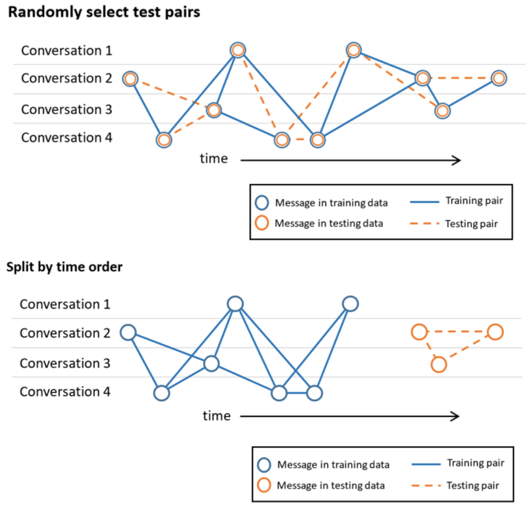


Figure 2: Split training data and testing data (a) randomly or (b) time order

split into training and testing sets as depicted in Figure 2(a). However, the performance of SHCNN model drops significantly when we use a time order splitting method as shown in Figure 2(b). In other words, the high performance reported in the paper may due to data peeping rather than a good model. In fact, since the messages in the same-thread could fork subtopics as the conversation goes on, it is difficult to judge whether two messages are related to each other directly, even for humans.

2.1 Message pair classification tasks

To build a better model for reply prediction tasks, we also refer to other tasks that accept two messages as input such as aspect-level sentiment analysis and natural language inference.

Aspect-level sentiment analysis aims to determine the sentiment polarity of a review sentence with respect to a given aspect. Many models and methods have been proposed from traditional machine learning methods (Schouten and Frasincar, 2016) to deep learning models (Zhou et al., 2019). For example, Wang et al. (Wang et al., 2016) proposed an attention-based LSTM network for aspect-level sentiment classification. Huang et al. (2018) introduced an attention-over-attention (AOA) neural network to capture the interaction between aspects and context sentences. The AOA model outperforms previous LSTM-based architectures. More models on aspect-level sentiment analysis can be found at (Zhou et al., 2019).

On the other hand, the task of natural language inference is to determine if one given statement (a premise) semantically entails another given statement (a hypothesis). For example, Parikh et al. (2016) proposed ”Decomposable Attention Model” which uses a shared sentence representation with fewer parameters and mutual attention mechanism to build a model with high performance.

2.2 Learning sentence representation

Note that most models mentioned above use pretrained word embedding for the input layer and adopt RNN or CNN to learn sentence representation. Recently, learning sentence representation from large unlabeled corpus has become feasible. For example, Radford et al. Radford2018ImprovingLU suggested a two-stage training process called Generative Pre-Training (GPT). Devlin et al. (2018) improved GPT with Bidirectional Encoder Representations from Transformers (BERT), which also uses two-stage training process and stacked Transformer. Two tasks are considered in the pre-training stage, including masked language model (MLM) and next sentence prediction (NSP). Many NLP tasks built on top of BERT have been shown to surpass the previous state-of-the-art systems, including question answering and sentiment analysis (Sun et al., 2019).

3 Problem Definition and Datasets

Both the **same-thread** and **direct-reply prediction** tasks are binary classification problems with training data represented as (x,y), where x=(m_1,m_2) and $y \in \{0,1\}$ denoting whether m_1 and m_2 come from the same topic or m_2 is a reply of m_1.

There are two data sets used in this paper, namely IRC (Internet Relay Chat) and Reddit.

The IRC data set is a manually tagged data set used in (Elsner and Charniak, 2008). IRC provides group chats thus multiple conversations are interspersed with each other in a channel. This data set contains 6 hours of messages from the LINUX channel. Each message is annotated with the conversation or thread it is involved. Thus, it is consistent with the same-thread task.

The Reddit Dataset consists of comments from

Dataset	Reddit			IRC
	Gadgets	Iphone	Politic	
Conversations	468	529	6,197	159
Messages	11,071	10,261	148,942	1,865
Speakers	6,387	4,506	28,365	183
All pairs	487,695	507,226	4,492,361	79,682
same-thread pairs	118,889	111,145	1,226,863	5,390

Table 1: Datasets for the same-thread task

Dataset	Direct-Reply Prediction Task		
	Gadgets	Iphone	Politic
Conversations	18,220	34,348	27,361
Messages	358,212	345,487	800,619
Speakers	118,225	49,571	72,787
All pairs	1,143,058	947,484	2,423,900
Reply pairs	228,438	189,353	477,780

Table 2: Datasets for the direct-reply prediction task

Reddit articles which are synthesized by following Jiang et al.'s work (Jiang et al., 2018). Reddit is a web content rating and discussion website, where members can submit contents such as links or news or text posts which are then voted up or down. Posts are organized by subjects into user-created subreddit. Theoretically, all comments from the same post can be treated as the same-thread messages. Jiang et al. mix comments from different articles around the same time to create conversation logs of multiple people chat for the same-thread tasks, as shown in Figure 3.

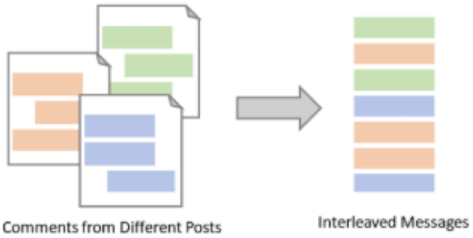


Figure 3: Combination of Reddit dataset

Since the original posts are usually longer than comments, we keep only messages that are replies to comments and remove comments to the original articles such that most messages are about the same length. We collect three subreddits from gadgets, iPhone and politics channels from 2016/06 to 2017/05 and remove articles with too many comments and enumerate message pairs that are within T (one-hour) time span to prepare training data for the same-thread task as stated by Jiang, et al.

We follow (Jiang et al., 2018) to synthesize the dataset: For the direct-reply prediction task, every comment in Reddit is a reply to a previous message, which makes it a good source for the direct-reply prediction task. For this task, each comment is paired with five messages: with only one correct reply message and four other messages within time interval T (=1 hour), which may come from the same-thread or different threads. Table 1 and Table 2 show the number of conversations, messages, average messages per conversation, speakers, message pairs prepared for the same-thread task and the direct-reply prediction task, respectively. Because random splitting of data into training, validation and testing data may cause the message pairs from the same-thread to occur in both training and testing sets as shown in Figure 2(a), making the performance untrustworthy, we split the data based in chronological order (Figure 2(b)) to avoid data peeping, and use the first 72% of data for training, the following 8% for validation and the last 20% for testing.

4 Methods

In this paper, we consider models based on GloVe word embedding and BERT models for message-pair classification.

4.1 Glove-based Representation

A typical neural network model consists of embedding layer for word representation, hidden layer such as mutual attention for message representation, and output layer for prediction. For embedding layer, we adopt pre-trained GloVe (Pennington et al., 2014) word embedding matrix from Common Crawl dataset (840B tokens), which contains a case-sensitive vocabulary of size 2.2 million. Given a message $m = [w_1, w_2, \ldots, w_L]$ with L tokens, we look up the embedding vector $u_i \in R^{d_w}$ for each word. If a token in the message does not exist in the pretrained model, we replace it with the UNKNOWN token embedding. Thus, the message m is represented by a $L \times d_w$ matrix $\mathbf{U} = [u_1, u_2, \ldots, u_L]$.

We consider two models for message representation. The first one is GCNN-LSTM, and the second is LSTM with dual attention.

GCNN-LSTM Representation

Convolutional neural networks (CNN) are shift invariant artificial neural networks with shared-weights architecture and translation invariance characteristics, commonly applied in image processing. With d_c kernels of size $d_w \times k$, the output of the 1-D CNN layer will be $L \times d_c$ feature matrix. Here, we use Gated Linear Unit (GLU) proposed in (Dauphin et al., 2017) to control which information flows in the network.

$$\tilde{\mathbf{U}} = (\mathbf{U} * W + \mathbf{b}) \otimes \sigma(\mathbf{U} * W' + \mathbf{b}') \quad (1)$$

where $W, W' \in R^{k \times d_w \times d_c}$ and $\mathbf{b}, \mathbf{b}' \in R^{d_c}$ denote the parameters for two CNNs, one for feature extraction and one for GLU.

To deal with word sequence, we adopt a BiLSTM layer to capture the message information. LSTMs (Long Short-Term Memory) are recurrent neural networks that are able to capture ordered information from input sequence of tokens. BiLSTM is obtained by stacking two LSTM networks to get information from backwards and forward states simultaneously.

$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i} \quad (2)$$

Let θ_r denotes the parameters for BiLSTM, we define the output of BiLSTM function $G_r(\tilde{U}; \theta_r)$ as the sum of all hidden states, i.e.

$$G_r(\tilde{U}; \theta_r) = h_1 + h_2 + \ldots + h_L \quad (3)$$

Prediction and Objective Function

Let v_1 and v_2 denotes the output of the two input messages m_1 and m_2. For prediction, we use two fully connected layers to make the prediction, i.e.:

$$\hat{y} = \sigma(ELU([v_1, v_2]^T W_0 + b_0) W_1 + b_1) \quad (4)$$

where $W_0 \in R^{2d_r \times d_f}$, $W_1 \in R^{d_f \times 1}$. Let Θ denote the parameters used in the model to encode the sentence, i.e. $\Theta = \{W, W', b, b', \theta_r, W_0, b_0, W_1, b_1\}$, the model is trained by minimizing cross-entropy with L2 regularization as shown below.

$$\begin{aligned} Loss(D) &= \textstyle\sum_{(x,y) \in D} y \cdot \log \hat{y} \\ &+ (1-y) \cdot \log(1-\hat{y}) + \lambda \left\|\Theta\right\|^2 \end{aligned} \quad (5)$$

LSTM-Dual Attention Model

Inspired by the power of attention mechanism, the second model we proposed is BiLSTM with dual attention.

Attention Mechanism is originally designed to help the decoder of seq2seq model generate words one by one. The idea is to collect the output vector h_i ($\in R^{2d_r}$) at each word for attention. Let $z_1 = G_r(\mathbf{U_1}; \theta_r)$ and $z_2 = G_r(\mathbf{U_2}; \theta_r)$, we can exploit attention mechanism to generate a representation for z_1 based on the content of z_2. Specifically, we calculate the weighted sum of the output at each step of the first message, $[h_1^1, h_2^1, \ldots, h_L^1]$ with weight vector α^1 as below:

$$z_1' = \sum_{i=1}^{L} \alpha_i^1 h_i^1 \quad (6)$$

where $\alpha^1 = [\alpha_1^1, \alpha_2^1, \ldots, \alpha_L^1]$ is computed by taking the inner product of z_2 with each h_i^1.

$$\alpha_i^1 = softmax(z_2^T h_i^1) = \frac{exp(z_2^T h_i^1)}{\sum_{j=1}^{L} exp(z_2^T h_j^1)} \quad (7)$$

Similarly, z_2' is calculated by the weighted sum with weight vector α^2. Finally, we concatenate z_1 with z_1' to form v_1 ($\in R^{4d_r}$), i.e. $v_1 = z_1 \oplus z_1'$, and z_2 with z_2', i.e. $v_2 = z_2 \oplus z_2'$, to form v_2 for classification task as described in Equation 4.

Attention over Attention (AOA) Model

For two output hidden states from BiLSTM $h_1 \in R^{n \times 2d_h}$ and $h_2 \in R^{m \times 2d_h}$, AOA (Huang et al., 2018) first calculates a pair-wise interaction matrix $I = h_1 \cdot h_2^T$, where the value of each entry I_{ij} represents the correlation of a word pair among the two input messages) and compute both column-wise softmax, $\alpha \in R^{n \times m}$ and row-wise softmax, $\beta \in R^{n \times m}$.

$$\alpha_{ij} = \frac{exp(I_{ij})}{\sum_i^n exp(I_{ij})}, \beta_{ij} = \frac{exp(I_{ij})}{\sum_j^m exp(I_{ij})}, \quad (8)$$

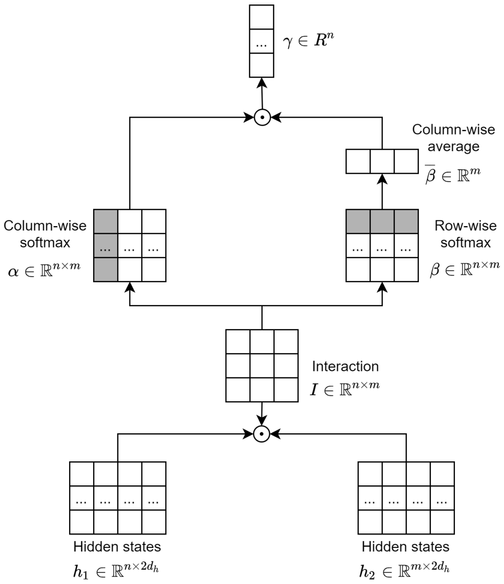

$\gamma \in R^n$
Column-wise average
$\overline{\beta} \in \mathbb{R}^m$
Column-wise softmax
$\alpha \in \mathbb{R}^{n\times m}$
Row-wise softmax
$\beta \in \mathbb{R}^{n\times m}$
Interaction
$I \in \mathbb{R}^{n\times m}$
Hidden states
$h_1 \in \mathbb{R}^{n\times 2d_h}$
Hidden states
$h_2 \in \mathbb{R}^{m\times 2d_h}$

Figure 4: An Attention-over-Attention Module (AOA) (Huang et al., 2018)

AOA Layer

The idea of AOA is to compute the attention weight over the averaged attention weight $\overline{\beta} \in \mathbb{R}^m$ where $\gamma \in \mathbb{R}^n$,

$$AOA(h_1, h_2) = \gamma = \alpha \cdot \overline{\beta}^{\mathsf{T}}. \tag{9}$$

where

$$\overline{\beta}_j = \frac{1}{n}\sum_i^n \beta_{ij}. \tag{10}$$

We call γ the output of AOA layer and use it to calculate the final sentence representation $r \in R^{2d_h}$.

$$r(h_1, h_2) = h_1^{\mathsf{T}} \cdot \gamma \tag{11}$$

The final sentence representation r is then used for final result prediction, i.e. $\mathbf{p_o} = \mathbf{r}$.

$$P(y|x) = \sigma(\mathbf{w} \cdot \mathbf{p_o} + b_o) \tag{12}$$

The attention-over-attention layer structure is as shown in Figure 4.

4.2 BERT Based Models

Different from context-free models, which generate a fixed word embedding representation for each word in the vocabulary, BERT is able to give a context-dependent representation of the words. Consequently, we use the BERT model released by Google and fine-tune it for the same-thread/direct-reply prediction task.

Given two input messages m_1 (with length n) and m_2 (with length m), we employ BERT component with L transformer layers to calculate the corresponding contextualized representations with input of the form $([CLS], m_1, [SEP], m_2)$. Let H^l be the output of the transformer at layer l, thus $H^i = [h_0^l, h_1^l ... h_{n+m+2}^l]$ can be calculated by

$$H^{i+1} = BiTransformer(H^i), \tag{13}$$

The basic BERT sentence pair classification (BERT-SPC) model takes the output of [CLS] token as the prediction layer input, i.e. $\mathbf{p_o} = \mathbf{H_0^L}$ by Eq 12. The entire model is fine-tuned with a standard cross-entropy loss with L2 regularization.

BERT-SPC-AOA Model

To further improve BERT-SPC model, we concatenate the output $r(h_1, h_2)$ with [CLS] output as the input to the prediction layer, i.e. $\mathbf{p_o} = r(h_1, h_2) \oplus H_0^L$ by Eq. 12.

5 Experiments and Analysis

For non-BERT deep learning models, the pre-trained word embedding is GloVe (Pennington et al., 2014) with case distinction trained on the Common Crawl dataset to distinguish English word embedding, dimension d_w is 300 and total 2.2 million words. The hidden layers in BiLSTM h_r are 128, the number of kernels used in CNN h_c is 128, and kernel size k is 5. All models are implemented with Tensorflow. The batch size used in the traditional deep learning model is 256 and the maximum epoch and initial learning rate are set to 40 and $2 * 10^{-4}$.

For BERT model, the batch size is 32. The maximum epoch and initial learning rate are 6 and $2{*}10^{-5}$, respectively. The optimizer used in all models is Adam with $\beta 1$ = 0.9 and $\beta 2$ = 0.999. The L2 weight λ of the objective function in Equation 4 is set to 0.01. We apply linear attenuation to the learning rate and use warmup in the first 30% of the training step with dropout set to 0.1. For training data, the ratio of the positive versus negative (different thread) pairs is 1:1.

Dataset		Reddit						IRC	
		Gadgets		Iphone		Politic			
Measure		F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
Rand. Split	SHCNN	0.779	0.887	0.608	0.816	0.639	0.770	0.805	0.967
	GCNN+LSTM	0.981	0.990	0.956	0.980	0.945	0.969	0.319	0.801
	LSTM+DualAtt	0.809	0.900	0.618	0.804	0.638	0.775	0.346	0.854
	AOA	0.979	0.990	0.933	0.970	0.812	0.887	0.571	0.915
	BERT SPC	**0.988**	**0.994**	**0.994**	**0.994**	**0.996**	**0.970**	**0.888**	**0.985**
Time Order Split	SHCNN	0.253	**0.709**	0.318	**0.553**	0.411	0.553	0.039	**0.939**
	GCNN+LSTM	0.289	0.645	0.364	0.516	0.309	0.601	0.098	0.630
	LSTM+DualAtt	0.308	0.710	0.417	0.500	0.441	0.616	0.089	0.854
	AOA	0.310	0.634	**0.428**	0.541	**0.482**	0.560	0.084	0.557
	BERT SPC	**0.522**	0.508	0.298	0.534	0.423	**0.667**	**0.223**	0.898

Table 3: Performance comparison of the same-thread prediction task.

Evaluation method

Because both tasks are modeled as binary classification problems and the numbers of same-thread pairs or direct-reply message pairs are fewer than different thread and non-reply message pairs, we consider them as positive data and use F1 and accuracy for the evaluation.

5.1 Same-thread Task

We start with the same-thread prediction task. Table 3 shows the performance of the models trained on a random or time order split data sets. As we can see, all the proposed models outperform SHCNN. GCNN-LSTM model and BERT SPC model perform especially good on random split data with 0.981 and 0.988 F1 on Reddit Gadgets dataset. However, the performance of all models drops significantly when data is split by time order. The F1 of the Reddit datasets nosedives from above 0.9 to 0.2 and 0.3 for GCNN+LSTM model. Even for BERT model, the plunge on IRC dataset is also precipitous (from 0.888 to 0.223).

In order to understand why all models perform poorly in the same-thread task, we conducted an error analysis and found the task to be challenging even for human beings. Table 4 shows some mislabeled message pairs and their ground truth labels. We notice that it is not easy to recognize the connections between the message pairs (e.g. the first three message pairs) that come from the same-thread without context. Meanwhile, annotators might be misled to give positive labels when two messages mention about the same entity, while the messages actually come from two different conversations. For example, the last message pairs both mentioned about Hillary, but the messages actually come from two threads. Thus, we argue that the same-thread task is a very difficult task when no context is given. However, context might not always help when multiple threads are mixed together as studied in (Wang and Oard, 2009).

To confirm our speculation, we randomly select 500 message pairs from Reddit test data, and give them to four graduate students to judge if each message pair comes from the same-thread. Only one out of 4 annotators is able to achieve higher than 0.5 F1 and the average F1 is only 0.340 (Table 6), indicating the difficulty of the same-thread task. In fact, it is not enough for annotators to rely on only two messages to determine whether they are from the same-thread. On the other hand, the performance could be greatly improved to 0.660 F1 and 0.883 accuracy in the direct-reply predication task.

5.2 Direct-reply Prediction Task

In view of the above problem, we consider the direct-reply prediction task and only split data based on time order to see how well models could perform for future application. As shown in Table 6, the performance of manual annotation on the direct-reply prediction task is much better than that for the same-thread task. Most models have performance higher than 0.5 F1 for the direct-reply prediction task.

As shown in Table 5, LSTM with dual attention

m_1	m_2	Label
He actually didn't shoot anyone who was walking down the street.	Agreed. A shallow grave in the woods is more fitting.	True
He teared up thinking about how this will make his re-election campaign more difficult.	House Freedom Caucus "You're free to die, you sick moochers. Isn't it beautiful!"	True
I wonder if he used all the best words?	They can't have my brand!	True
Trump tries to clean up on Whitewash Crimea	Trump is a traitor and will sell this country out to the highest bidder the moment he gets into office.	False
Pepperidge Farms remembers people saying Hillary "Warmonger" Clinton.	I love how he framed Hillary as the warhawk.	False

Table 4: Some mislabeled examples and their true labels for the same-thread task.

direct-reply Task	Gadgets		Iphone		Politic	
	F1	Acc	F1	Acc	F1	Acc
SHCNN	0.513	0.845	0.455	0.849	0.512	0.849
GCNN+LSTM	0.591	0.856	0.534	0.845	0.573	0.861
LSTM+DualAtt	0.598	0.869	0.567	0.856	**0.637**	**0.870**
AOA	0.514	0.721	0.486	0.723	0.566	0.778
BERT SPC	0.632	**0.883**	0.543	**0.862**	0.616	0.818
BERT-SPC-AOA	**0.669**	0.877	**0.617**	0.840	0.623	0.825

Table 5: Model performance for the direct-reply prediction task

	same-thread		direct-reply Prediction	
Measure	F1	Acc	F1	Acc
Annotator 1	0.105	0.728	0.548	0.860
Annotator 2	0.352	0.763	0.800	0.930
Annotator 3	0.263	0.732	0.590	0.875
Annotator 4	0.638	0.796	0.703	0.865
average	0.340	0.755	0.660	0.883

Table 6: Performance of four annotators for the same-thread and direct-reply prediction task.

outperforms BERT on two of the Reddit datasets except for Gadgets. However, the best performance on Gadget Reddit dataset is achieved by BERT-SPC-AOA model. Compared with the best result (0.522, 0.428 and 0.482 F1) for the same-thread prediction task, We can see the performance on three Reddit datasets is improved to 0.669, 0.617, and 0.637 F1 by BERT-SPC-AOA and LSTM with dual attention model.

6 Conclusion

This paper addresses the rationality of the chatlog disentanglement problem in two ways. First, the testing data should be prepared to simulate future unseen data. Second, the same-thread task without context information is too challenging even for human annotators. Thus, we propose the direct-reply prediction task for question-answer pair generation from chatlogs. In the direct-reply prediction task, using the pre-trained BERT model for Fine-Tuning can achieve good performance, even with less training data. However, the data quality used by the downstream task seems to be a big problem when using BERT for Fine-Tuning. When a large number of messages are disentangled, the negative examples for both the same-thread or direct-reply prediction tasks are much larger than the positive examples. Though, down sampling is adopted to balance the training data, the models still tend to classify the message pairs to the negative session, leading to low F1.

For future work, how to add context information is an alternative direction to consider. Meanwhile, as direct-reply predication is not a symmetric problem, an input of (m_1, m_2) is different from (m_2, m_1). Thus, we wonder that the shared message representation and the predication function of multi-layer perception might not be enough to extract the features from two message representation. Thus, we might consider asymmetric models, i.e. two message encoders for each of the input messages to see if it could achieve better performance.

Acknowledgments

The research is partially supported by Ministry of Science and Technology, Taiwan under grant MOST109-2221-E-008-060-MY3.

References

James Allan, 2002. *Introduction to Topic Detection and Tracking*, pages 1–16. Springer, Boston, MA.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio, June. Association for Computational Linguistics.

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. *CoRR*, abs/1804.06536.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September. Association for Computational Linguistics.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

K. Schouten and F. Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message streams. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 35–42, New York, NY, USA. Association for Computing Machinery.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Lidan Wang and Douglas W. Oard. 2009. Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, page 200–208, USA. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Jie Zhou, Jimmy Xiangji Huang, Qin Chen, Qinmin Vivian Hu, Tingting Wang, and Liang He. 2019. Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. *IEEE Access*, 7:78454–78483.

Japanese Named Entity Recognition from Automatic Speech Recognition Using Pre-trained Models

Seiichiro Kondo[1] **Naoya Ueda**[1] **Teruaki Oka**[1]
Masakazu Sugiyama[2] **Asahi Hentona**[2] **Mamoru Komachi**[1]
[1]Tokyo Metropolitan University [2] AI Shift
{kondo-seiichiro, ueda-naoya}@ed.tmu.ac.jp, teruaki-oka@tmu.ac.jp
{sugiyama_masakazu, hentona_asahi}@cyberagent.co.jp,
komachi@tmu.ac.jp

Abstract

Japanese named entities extracted from automatic speech recognition frequently contain speech recognition errors and unknown named entities due to abbreviations and aliases. One possible solution to this problem of the named entity extraction task is to use a pre-trained model trained on a large quantity of text to acquire various contextual information. In this study, we performed named entity recognition on the logs of a task-oriented dialogue system for road traffic information in Fukui, Japan, using pre-trained BERT-based models and T5. In our experiments using our prepared data, the F1 scores of BERT and T5 are higher than that of string match by 20.2 point and 21.1 points, respectively. The results confirmed that these pre-trained models exhibited significantly higher accuracies on unseen entities than methods based on dictionary matching.

1 Introduction

In interactive voice response services, it is necessary to accurately perform entity extraction and mention detection from the speech recognition results of user's speech to carry out dialogue via voice. However, this input contains a variety of noise, including speech recognition errors, compared to written text. Moreover, users' utterances may include personal representations for a certain entities. In our full research, we plan to address these issues in two steps: named entity recognition (NER) and entity linking (EL). The first step, NER, recognizes a part of the text that is likely to represent a named entity. In the second step, we then attempt to link it to a prepared set of entities even if it contains errors. With this architecture in mind, in this study, we tackle the first step, which is NER from speech recognition results.

This study focused on named entity recognition (NER) in the context of a task-oriented dialogue system that provides information in response to the user's requests pertaining to road traffic. In our system, NER is accomplished by linking the automatic speech recognition (ASR) text with a dictionary created for each task (see Fig. 1). One way to achieve more accurate recognition is to extract the named entities before the linking.

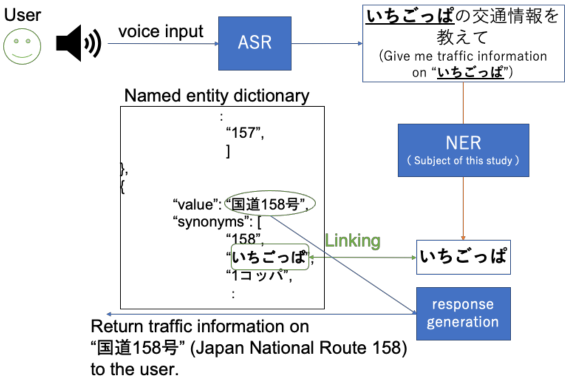

User
voice input
ASR
いちごっぱの交通情報を教えて
(Give me traffic information on "いちごっぱ")
NER
(Subject of this study)
Named entity dictionary
:
"157",
]
},
{
"value": "国道158号",
"synonyms": [
"158",
"いちごっぱ",
"1コッパ",
:
Linking
いちごっぱ
response generation
Return traffic information on "国道158号" (Japan National Route 158) to the user.

Figure 1: Flowchart of our envisioned spoken dialogue system. This study focuses on the NER component. The goal is to extract named entities even from noisy text caused by speech recognition errors or abbreviations.

Although we enforce the inputs to include location names by system-driven conversation, speech recognition errors and unknown named entities from abbreviations and aliases may occur. For example, "いちごっぱ" (*ichi-go-ppa*, 158) is a colloquial expression for "国道158号" (*kokudou-hyaku-goju-hachi-gou*, Japan National Route 158). In addition, there are many cases in Japanese where the sounds are close, but the surface forms and meanings are entirely different. Therefore, in this setting, NER using conventional methods is difficult, especially for rule-based methods.

To improve text processing functionality, this study focused on NER on ASR texts. [1] We focused on context-based NER in the ASR texts because named entities may be unknown surfaces as

[1] Note that, although Omachi et al. (2021) postulated that an end-to-end (E2E) approach for processing speech recognition results might be preferable, we used existing ASR to enable the flexible exchange of modules and resources, making it necessary to process ASR texts.

a result of speech recognition errors or abbreviations. Based on the assumption that contextual information can be used effectively by pre-trained models trained on a large number of sentences, we used BERT-based large-scale pre-trained models for NER (Devlin et al., 2019; Clark et al., 2020). We also investigated the performance of T5 (Raffel et al., 2020), a pre-trained encoder-decoder model.

2 Related Work

NER from ASR. Wang et al. (2021) performed NER from speech recognition using the matching approach. They used the embeddings of the top N prediction candidates of ASR. In this study, we experimented with only the top predicted ASR candidate, and performed string matching with rule-based NER for simplicity.

NER from speech recognition with neural models for English has been studied previously. Raghuvanshi et al. (2019) extracted personal names from text containing speech recognition errors using additional information not contained in the text and reported that the recall was improved. Yadav et al. (2020) studied the E2E approach and were able to extract named entities robustly and efficiently. We used neural models to perform NER from Japanese ASR texts under the assumption that the ASR architecture cannot be changed. In other words, in this study, experiments will be conducted in a setting where only the text processing section is involved and no other information, such as voice information, is used.

Japanese NER. Rule-based matching methods (Sekine et al., 1998) and machine learning-based methods (Utsuro and Sassano, 2000; Sassano and Utsuro, 2000) have been proposed for Japanese NER. However, these studies focused on manually written texts, whereas ASR texts often contain problems specific to speech recognition, such as speech recognition errors. In this study, we attempted to extract named entities from such ASR texts.

NER in Japanese speech recognition has been performed using support vector machines (SVMs). Sudoh et al. (2006b,a) reported that when training SVMs on ASR text, precision can be improved by incorporating a confidence feature that indicates whether a word is correctly recognized. In contrast, we aimed to extract named entities from text containing speech recognition errors, focusing on recall to lead to subsequent linking tasks and using a pre-trained model for this purpose.

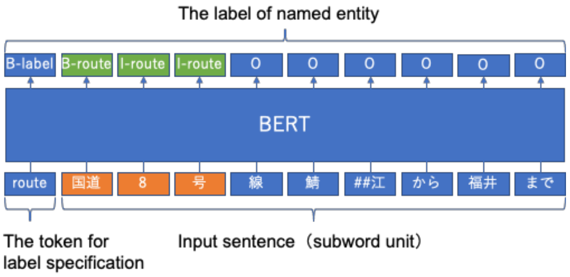

The label of named entity
B-label
B-route
I-route
I-route
O
O
O
O
O
O
BERT
route
国道
8
号
線
鯖
##江
から
福井
まで
The token for label specification
Input sentence（subword unit）

Figure 2: NER using BERT-based models

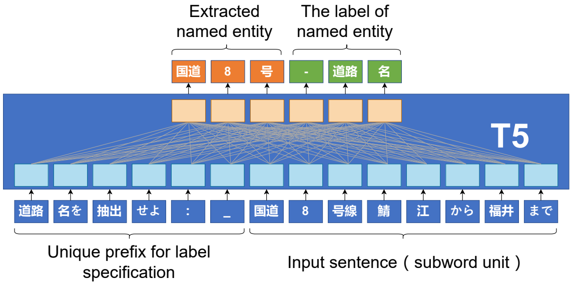

Extracted named entity
The label of named entity
国道
8
号
-
道路
名
T5
道路
名を
抽出
せよ
:
_
国道
8
号線
鯖
江
から
福井
まで
Unique prefix for label specification
Input sentence (subword unit)

Figure 3: NER using T5

3 Our Method

In this study, we used road traffic data for NER evaluation of ASR text containing speech recognition errors and obtained named entities related to roads and addresses. We assumed that the output labels (roads and addresses) could be specified as a precondition.

3.1 NER using BERT based models

Devlin et al. (2019) demonstrated the state-of-the-art performance of a fine-tuned BERT model on the CoNLL-2003 NER task (Tjong Kim Sang and De Meulder, 2003). Following their approach, we considered NER as a sequence labeling task. The text was tokenized, split into subwords, and labeled based on the BIO model, in which "B" was assigned to the beginning, "I" was assigned to the interior and the end of the named entities, and "O" was assigned to any other tokens. The schematic view is presented in Figure 2. Labels for road information were considered as "`{B, I}-route`", and labels for address information as "`{B, I}-address`". To specify a label, we prefixed the statement with a "route" or "address" token and gave "B-label".

	text
match	鯖江から敦賀市へ向かう高速道路 (Highway from Sabae to Tsuruga City)
fallback	えーとサザエさん、サザエ市春江町 (Well, Sazae-san, Sazae City, Harue-cho)

Table 1: Example of match and fallback data (the underlined parts are named entities):サザエ (*sazae*, turban shell) is a recognition error of 鯖江 (*sabae*), which is the name of a city in Fukui.

		train	dev	test
match	utterance	1,757	220	220
	address	1,220	144	147
	route	802	104	110
fallback	utterance	949	118	122
	address	197	30	26
	route	92	8	17

Table 2: The number of data instances used in the experiment (the number of utterances and named entities with each label).

3.2 NER using T5

We performed NER with a Seq2Seq pre-trained model because Constantin et al. (2019) reported that Seq2Seq models achieve excellent sequence labeling of noisy texts. Also,Phan et al. (2021) performed NER using domain-adapted T5 on medical literature. Their experimental results show that NER can be performed with high F1 scores even with a seq2seq model, namely T5. Following their approach, we considered NER as a question-and-answer task. A text with a special label at the beginning was the input sequence and named entities corresponding to this special label were output. The system was set up such that each extracted named entity was added to the end with a label followed by a dash. The schematic view is presented in Figure 3. Labels for road information were considered as "道路名" (road name), and the labels for address information as "住所名" (address name). To specify the label, the special tokens "道路名を抽出せよ:" (extract road names) or "住所名を抽出せよ:" (extract address names) were added at the beginning of the sentence.

4 Experiments

4.1 Data

In this study, we conducted NER using a system-driven dialogue log containing road traffic information in Fukui, Japan [2]. The dialogue logs were obtained from the turns where the user seemed to have uttered the names of roads or addresses based on the conversation before and after. A dictionary of the named entities to be extracted was provided. In this dictionary, aliases, abbreviations, and speech recognition errors were registered (the dictionary is shown in Fig. 1). The target texts for which NER succeeded and failed were *match* and *fallback*, respectively; an example is presented in Table 1.

The match data were labeled by dictionary matching, with incorrect labels manually removed. For the data in fallback, we manually annotated named entities related to the road and the address. This annotation was performed considering speech recognition errors and any named entities existing in Fukui even though not in the dictionary. Notably, because fallback data were annotated based on whether the named entities exist in Fukui, a difference existed in the criteria of labeled words between match and fallback data. We randomly split match and fallback data so that the training, development, and test data were 8:1:1. The number of data instances is shown in Table 2. Note that these data were not arbitrarily sampled but were obtained over a certain period.

Achieving accurate NER with match and small fallback training data is practical, since the former requires only a reasonably sized dictionary but the latter needs human annotation. For data collection, we considered match data as inexpensive to obtain because they could be extracted by dictionary matching, and fallback data as expensive because they could not (Subsection 4.4).

4.2 Setting

We compared four NER systems, viz., a string-matching model based on a dictionary, two pre-trained BERT-based models, and T5. The script of transformers, published by Huggingface [3], was used for fine-tuning all models.

For the BERT-based models, we used the BERT

[2] We will perform NER on other dialogue log data as well in the future.

[3] https://github.com/huggingface/transformers

method	data	P	R	F1	c_P	c_R	c_F1	P	R	F1	c_P	c_R	c_F1
String Match	M	96.3	**100**	**98.1**	96.3	**100**	98.1	—	—	—	—	—	—
	F	50.0	23.3	31.7	50.0	23.3	31.7	—	—	—	—	—	—
		trained using all data						trained by match data					
BERT	M	97.3	97.3	97.3	**99.2**	99.2	**99.2**	97.3	97.3	97.3	98.8	98.8	98.8
	F	67.9	83.7	75.0	67.9	83.7	75.0	**58.8**	46.5	**51.9**	**58.8**	46.5	**51.9**
ELECTRA	M	96.9	98.1	97.5	98.1	99.2	98.6	**97.7**	**98.1**	**97.9**	**99.2**	**99.6**	**99.4**
	F	66.0	72.1	68.9	66.0	72.1	68.9	54.5	41.9	47.4	57.6	44.2	50.0
T5	M	**98.0**	97.7	97.9	**99.2**	98.8	99.0	97.3	97.7	97.5	98.5	98.8	98.6
	F	**74.0**	**86.0**	**79.6**	**74.0**	**86.0**	**79.6**	41.3	**60.5**	49.1	42.3	**62.8**	50.9

Table 3: Experimental results for string matching, BERT, ELECTRA, and T5. M and F denote match and fallback data, respectively. "c_" means results of re-scoring named entities as true positives when they are predicted to be longer than the reference.

published by Tohoku University [4] and ELECTRA published by Megagon Labs [5]. We fine-tuned both models through token classification. For BERT, we set the batch size to 8, the number of epochs to 3, and the maximum sequence length to 258. For ELECTRA, we set the learning rate to 0.00005, the batch size to 8, the number of epochs to 20, and the maximum sequence length to 128. The T5 model was fine-tuned from the model published by Megagon Labs [6]. We set the learning rate of T5 to 0.0005, the batch size to 8, the number of epochs to 20, and the maximum sequence length to 128. Note that the pre-training datasets for T5 and ELECTRA were approximately the same size, while that for BERT was much smaller.

4.3 Evaluation

We evaluated the performance by calculating precision, recall, and F1 scores, considering a perfect match as a true positive. Because of the difference between the labeling criteria of the match and fallback data (Subsection 4.1), we evaluated each test set separately. Evaluation of match data serves as a measure of whether named entities flagged by dictionary matching can be extracted, whereas evaluation of fallback data measures whether it is possible to extract named entities that are not included in the dictionary.

In this study, we assumed that entity linking was performed in the downstream task. If the extracted named entities are shorter than the original entities, linking may become problematic. In contrast, when the extracted named entities are longer than the original entities, the problem in linking is considered minor. Therefore, under the lenient evaluation setting, we considered the cases in which the named entities were covered as true positives, but we still considered the cases in which the partial matches were not covered as false positives. For example, if the named entity "**8号線**" (Route 8) is in the reference, the extraction of "**国道8号線**" (Japan National Route 8) is acceptable, but the extraction of only "**8号**" (Route 8) is not acceptable.

4.4 Results

Experimental results for the match and fallback test sets are presented in Table 3 for when both datasets were used as training data and for when only the match data was used.

String Matching For the match data, the recall was 100 because the data was created using dictionary matching. Precision was not 100 because we manually removed mislabeled data (Subsection 4.1) when creating the match data. Conversely, for the fallback data all the evaluation scores were less than 50.0, and so the unseen named entities were not sufficiently extracted.

BERT For the match test data, the F1 and c_F1 scores of BERT were comparable or superior to that of string matching, which was a desirable result for NER for subsequent tasks. For the fallback test data, the score was 20.2 points higher when training using only the match data and 43.3 points higher

[4] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[5] https://huggingface.co/megagonlabs/transformers-ud-japanese-electra-base-discriminator

[6] https://huggingface.co/megagonlabs/t5-base-japanese-web

model	text	translation
BERT (address)	吉田郡 **永平寺町**	(Yoshida-gun Eiheiji-cho)
T5 (address)	田尻町から**福井市**までの**福井市**内まで	(From Tajiri-cho to Fukui City to Fukui City)
BERT/T5 (route)	**イチゴったー**	(Ichigotta)

Table 4: Example of NER failure in match data. Bold and underlined texts denote the reference and hypothesis.

model	text	translation
BERT/T5 (address)	**横倉**ってどこや	(Where is Yokokura)
BERT (route)	**青年の道**	(Youth Road)
T5 (address)	低い	(low)
BERT/T5 (route)	**アイワかどう**	(Aiwakado)
BERT (address)	あの高みの方のエルパ行きのバスは取った後	(After taking the bus to Elpa at that height)
T5 (address)	あの高みの方のエルパ行きのバスは取った後	(After taking the bus to Elpa at that height)

Table 5: Example of NER in fallback data. Bold and underlined texts denote the reference and hypothesis.

when training using all data compared with string matching. In particular, the improvement in recall was remarkable, which indicated that BERT could extract unique named entities that could not be extracted by string matching. Adding the fallback data to the match data for training considerably increased the score. This increase is attributed to words not included in the match data (dictionary) being considered during training.

ELECTRA The score of ELECTRA was lower than that of BERT except for match test data when the model was trained using match data. This result shows that the amount of data used for pre-training has a small impact on the results of NER.

T5 The trend observed for T5 is the same as that for BERT. For the match test data, the performance was comparable to BERT. For the fallback test data, the precision was lower than that of BERT when training with only the match data. However, adding the fallback data to the match data for training improved the precision, and the F1 score was +4.6 points compared with BERT, which indicates that the extraction is consistent with the intention.

5 Discussion

Comparison to human performance To evaluate the upper limit of the fallback test data, we calculated the human recognition score by asking another person, not the annotator. Note that in human recognition, labeling is performed while checking whether the named entity exists in Fukui, considering the speech recognition errors. Precision, recall, and F1 were 80.0, 97.6, and 87.9, respectively. These results and Table 3 suggest that there is still room for improvement based on the performance of the pre-trained models.

method	error	false positive		false negative		
		NT	PM	ND	AE	others
human	11	8	2	0	0	1
BERT	21	14	3	2	2	0
T5	19	13	1	1	4	0

Table 6: Error analysis of each NER method in fallback data. We categorized the type of NER errors and the type of named entities that could not be extracted. NT: not tagged as named entity in test data. PM: partial match to the extraction span. ND: named entity not in the dictionary. AE: ASR error. Note that the number of samples in the NT and PM include those incorrectly labeled beginning with "I" by BERT.

Error analysis in fallback data Table 6 shows the classification results of NER error patterns for human and pre-trained models and the number of samples per type of named entities that could not be extracted in the fallback data. The number of extraction errors of BERT is higher than the others but is similar to that of T5.

In BERT outputs, some incorrect labels start with "I" for spans that might be part of named entities. These examples can be attributed to BERT's failed attempts to consider the context. Note that false positive extraction errors can be tolerated because entity linking is assumed to be performed subsequently in this setting.

Focusing on false negative errors, it seems that there is room for improvement in the extraction of named entities that are not included in the dictionary and named entities with speech recognition errors since a human could distinguish them. Since many of the error cases in the results obtained in this study are short in user speech, it may be necessary to use external data.

Examples Table 4 presents an example in which T5 successfully extracts but BERT fails, an example in which BERT successfully extracts but T5 fails, and an example in which both fail for match data. For fallback data, in addition to the same types of examples as Table 4, Table 5 shows an example of successful extractions with BERT and T5.

Because the test data of "match" are based on a dictionary match, the reference does not include "吉田郡" (*Yoshida-gun*) and "田尻町" (*Tajiri-cho*), but it must be noted that these are place names that exist in Fukui, and are therefore examples that should be extracted.

"イチゴったー" (*Ichigotta*) is thought to be a misrecognition of "いちごっぱ" (*ichi-go-ppa*, 158), which is sometimes uttered for "158号線" (*Hyakugojuhachi-gosen*, Route 158). Moreover, "アイワかどう" (*Aiwakado*) is thought to be a misrecognition of "舞若道" (*Maiwakado*), which is sometimes uttered for "舞鶴若狭自動車道" (*Maizuru-wakasa-jidosyado*). It is thought that the NER may be complicated by three main problems: simplified spoken language such as aliases or abbreviations, speech recognition errors, and specific to the Japanese language of notation, *kanji*, *hiragana*, and *katakana*. Currently, named entities such as these examples are handled using dictionaries, but the creation of dictionaries is time-consuming and they are limited in coverage. We plan to devise specific solutions for each of these problems in future work.

BERT and T5 can extract "横倉" (*Yokokura*). This named entity was not included in the training data even after adding fallback to the training data. This suggests that BERT and T5 can extract unknown named entities based on contextual information. Moreover, although "青年の道" (Youth Road) displayed in the fallback is not included in the training data, it is a road name that exists in Fukui. T5 was able to extract it because it predicted the road name from the word "道" (road) at the end. Only T5 could predict the road name from such a context, probably because of its different model structure and NER method. The identification of these factors is a subject for future research. Both BERT and T5 extracted "高みの（方の）" (height) as a named entity representing an address, which reveals that these models contextually tried to extract named entities from expressions that represent directions ("方").

"低い" (*hikui*, low) was extracted by T5. This may be a speech recognition error for "Fukui." Some of the user input in this experiment is shorter than a typical question answering task, and contextual information is not available in such examples. Nevertheless, the fact that T5 was able to extract this word is a remarkable result.

6 Conclusion

We performed Japanese NER on speech recognition using pre-trained BERT-based models and T5. The results of the experiment showed that data generated by dictionary matching was generally well extracted by the pre-trained models. Furthermore, by adding manually annotated data to the training data, we confirmed that it is possible to extract named entities not included in the dictionary. In future, we will consider more context-sensitive methods, including fine-tuning methods, to robustly extract named entities from noisy text containing unknown named entities, such as adding data that masks named entities to the training data.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*.

Stefan Constantin, Jan Niehues, and Alex Waibel. 2019. Incremental processing of noisy user utterances in the spoken language understanding task. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 265–274.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Motoi Omachi, Yuya Fujita, Shinji Watanabe, and Matthew Wiesner. 2021. End-to-end ASR to jointly predict transcriptions and linguistic annotations. In

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1861–1871.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Arushi Raghuvanshi, Vijay Ramakrishnan, Varsha Embar, Lucien Carroll, and Karthik Raghunathan. 2019. Entity resolution for noisy ASR transcripts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 61–66.

Manabu Sassano and Takehito Utsuro. 2000. Named entity chunking techniques in supervised learning for Japanese named entity recognition. In *Proceedings of the 18th Conference on Computatinal Linguistics (COLING)*, page 705–711.

Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. 1998. A decision tree method for finding and classifying names in Japanese texts. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 148–152.

Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. 2006a. Discriminative named entity recognition of speech data using speech recognition confidence. In *Proceedings of InterSpeech*, pages 337–340.

Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. 2006b. Incorporating speech recognition confidence into discriminative named entity recognition of speech data. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 617–624.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Takehito Utsuro and Manabu Sassano. 2000. Minimally supervised Japanese named entity recognition: Resources and evaluation. In *Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC)*, pages 1229–1236.

Haoyu Wang, John Chen, Majid Laali, Kevin Durda, Jeff King, William Campbell, and Yang Liu. 2021. Leveraging ASR N-Best in Deep Entity Retrieval. In *Proceedings of the Interspeech 2021*, pages 261–265.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from English speech. In *Proceedings of the Interspeech 2020*, pages 4268–4272.

Improving Automatic Evaluation of Acceptability Based on Language Models with a Coarse Sentence Representation

Vijay Daultani
Tokyo Institute of Technology
Meguro City, Tokyo
vijay.daultani@nlp.c.titech.ac.jp

Naoaki Okazaki
Tokyo Institute of Technology
Meguro City, Tokyo
okazaki@c.titech.ac.jp

Abstract

Motivated by recent findings on the probabilistic modeling of the acceptability judgments, several metrics have been proposed for its automatic evaluation. Frequently used metrics such as syntactic log odds ratio (SLOR) and its variants are based on utilizing probability from language model (LM) as a proxy for automatic acceptability evaluation of the generated text from an LM. Since one cannot use probability directly as a measure of acceptability, these metrics take steps to remove the confounding effects of noise from sentence length and lexical frequency to enable the usage of probability for acceptability evaluation. In this work, we argue that even though the effects are reduced, they still exist. We propose a data transformation strategy, Replace Named Entity (RNE), to get a coarse representation of a sentence to mitigate the remaining problems from lexical frequency. In RNE, we identify all proper nouns (i.e., NEs) in a sentence and classify them into one of eighteen types. Later RNE replaces all occurrences of NEs in a sentence with their identified type. We later trained three LMs (2, 3, 4-grams) and assessed their performance of five acceptability measures on four test datasets. We found that LMs trained on datasets preprocessed by RNE yield a significantly higher correlation (upto 52% on some datasets) with human acceptability judgment.

1 Introduction

Language Models (LMs) are often used to generate natural language text for NLP tasks — Machine Translation, Summarization, Question Answering, and many others. Moreover, intrinsic evaluation of the LMs often includes at least two characteristics (Mutton et al., 2007). First, how well the generated text represents the source data, whether it be the text in another language for machine translation, text to represent a summary of a document, or text to represent answers for a question, etc. Second, how well it conforms to regular human language use, a property we will refer to as acceptability of the sentence. Acceptability evaluation of a sentence is an essential task for automatically evaluating the quality of the text (to help filter unacceptable sentences) generated by the LMs.

Before moving forward, it is also essential to understand the difference between the usage of related words, i.e., fluency, readability, and grammaticality. Both fluency and readability are alternate words for acceptability, but the exact definition of these terms varies across the literature (Lau et al., 2017; Mutton et al., 2007; Kann et al., 2018; Storch, 2009; Pitler and Nenkova, 2008; Vadlapudi and Katragadda, 2010). However, we would like to differentiate acceptability from grammaticality. When a human evaluates the acceptability of a text, grammaticality is one of the possible factors, among others like semantic plausibility, processing difficulties, etc., that can also influence the acceptability of a given text. Though both 'acceptability' and 'grammaticality' have been used interchangeably, a sentence can be grammatical yet unacceptable and vice versa. A famous example is Chomsky's phrase, "Colorless green ideas sleep furiously." (Chomsky, 1957). Vice versa, acceptable sentences can be ungrammatical, e.g., in an informal context such as poems.

Sentence Type	Sentence
Original	Apple is set to hold its first event on Tuesday.
NER Result	[$_{\text{ORG}}$ Apple] is set to hold its [$_{\text{ORDINAL}}$ first] event of [$_{\text{DATE}}$ the year] on [$_{\text{DATE}}$ Tuesday].
Transformed	ORG is set to hold its ORDINAL event of DATE on DATE.

Table 1: Example sentence for motivation

Whether humans represent text acceptability evaluation as a binary classification (Warstadt et al., 2020) of acceptable vs. unacceptable class of sentence or as a probabilistic property (Lau et al., 2017) has been a subject of lengthy debate among cognitive scientists and linguists (Chomsky, 1957; Manning, 2002; Sprouse, 2007). Both of the above views have their strengths and weaknesses. On the one hand, binary classification models do not have the flexibility to distinguish text between varying degrees of acceptability. On the other hand, the acceptability of a sentence is not the same as the likelihood of its occurrence as determined by the probabilistic model, which depends on sentence length and lexical frequency. However, Lau et al. (2017) demonstrated it is possible to augment the probabilistic model to predict the acceptability of a sentence if one can normalize probability values from the LM to eradicate the confounding effects of noise introduced by length and lexical frequency, e.g., SLOR (Lau et al., 2017) and WP-SLOR (Kann et al., 2018).

In this article, we lean towards the view that acceptability is a probabilistic property. We depict that probability-based acceptability metrics SLOR and other variants, though they reduce the confounding effect of lexical frequency, do not resolve the problem entirely in Section 2. We later provide evidence that one reason for this problem is a granular-level representation of the sentence since LM has to predict the probabilities for all words in the sentence, including words for which LM has low confidence (i.e., rare words or out of vocabulary words). This work is motivated by how humans visualize the sentence (coarse-level representation) for acceptability evaluation. Our goal is to find how to generate such a sentence representation to enable LMs better correlate with human acceptability judgment for the existing metrics.

Table 1 presents a concrete example of our intuition and data transformation strategy. Original refers to the unprocessed sentence (granular-level representation). We consider replacing proper nouns, i.e., NEs, in a sentence to generate a coarse-level representation. We propose a two-step approach i.e., Replace Named Entity (RNE) (Section 3) to construct such a coarse-level representation. First, we employ Named Entity Recognition (NER), a task to identify the spans of text that constitute proper nouns in a sentence and classify each identified span into one of the NE types (i.e., subscript in the prefix) as shown in NER Result. Second, we replace each occurrence of a NE with its classified type within a sentence to generate the coarse-level representation (i.e., Transformed). We argue that the coarse-level representation closely resembles how humans will judge the Original sentence's acceptability. Therefore, it should be used as an input to train and test LMs. Our contributions are summarized as follows:

- We provide evidence that a popular probability-based acceptability evaluation metric SLOR has limitations since it is based on the lexical frequency of words in the training corpus.
- We demonstrate that the original sentence is not the best representation for training LM and propose RNE, a data transformation strategy to transform the sentences to a coarse-level representation.
- We present empirical evidence from our experiments that SLOR and other probability-based acceptability metrics correlate better with human judgment when LMs are trained on data processed with RNE.

To the best of our knowledge, ours is the first successful attempt to use NEs to find a coarse-level representation of a sentence that better resembles how humans evaluate acceptability to improve the correlation between existing automatic acceptability metrics and human acceptability judgment. In this paper, our target language is English. However, we believe the ideas and methods apply to other languages. Additionally, our proposed data transformation strategy is independent of both LM and the metric used to measure the acceptability.

2 The Problem

2.1 Problem Definition

Formally, a sentence S comprises of n words w_1, w_2, w_3, .. w_n. Each word w_i occurs with lexical frequency (count) f_i in the training corpus. The goal here is to find the acceptability $y \in \mathbf{R}_{\geq 0}$ of the sentence S.

2.2 Background

One might suggest treating the likelihood (probability) of a sentence S as its measure of acceptability, with 1 indicating completely acceptable and 0 means unacceptable sentence. Although, the idea seems enticing but will be an incorrect usage of the values in a probability distribution. The probability of a sentence, S, from an LM is the probability that a randomly selected sentence will be S and not a measure of its acceptability. Based on this observation, Lau et al. (2017) proposed several sentence and word level metrics to augment the probabilistic LM with an acceptability measure.

Among all proposed metrics, syntactic log odds ratio (SLOR) in Equation 1 has shown a good correlation with human acceptability judgment. SLOR is a function that normalizes the sentence probability and believed to eliminate the confounding factors of sentence length by dividing with the sentence length, i.e., $|S|$ and lexical frequency by subtracting the unigram probability of words comprising S. In Equation 1, $p_m(S)$ refers to the sentence probability of S, i.e., a product of probabilities assigned to each n-gram by the LM. $p_u(S)$ is the unigram probability for sentence S, i.e., a product of the unigram probabilities of the words comprised in the sentence.

$$\mathrm{SLOR}(S) = \frac{\log p_m(S) - \log p_u(S)}{|S|} \tag{1}$$

Against the expectation, we have observed that issues related to lexical frequency still persist in the formulation of SLOR. Essentially words lexical frequency in the training corpus can severely impact both sentence probability $p_m(S)$ and unigram probability $p_u(S)$, and therefore impairing the usage of SLOR for acceptability prediction.

To understand the impact of lexical frequency on SLOR, let's refer to three sentences, i.e., s1, s2, s3 in Table 2 with words France, Tuvalu, and Kiribati, referring to the names of three nations respectively. For our convenience, we will override the notation of $p_u(w)$ to refer to the unigram probability of the word w. To explain the issue, we have made two assumptions; first, let's assume that France occurs often and Tuvalu is a rare word in the training corpus. Second, let's assume the word 'Kiribati' never appears in the training corpus and is an out-of-vocabulary (OOV) word for the LM.

Index	Sentence
s1	He is a citizen of France.
s2	He is a citizen of Tuvalu.
s3	He is a citizen of Kiribati.

Table 2: Three sentences of equal length and equally acceptable

2.3 Unigram Probability

Based on our assumption about the frequencies of the word 'France' and 'Tuvalu', it will be safe to expect unigram probability $p_u(\text{France})$ to be higher than $p_u(\text{Tuvalu})$. In practice, to avoid the problem of 0 unigram probabilities with OOV words ('Kiribati'), it is common to replace them with UNK tokens in both training and test corpus and add UNK to the vocabulary. This will assign a tiny non zero unigram probability and therefore $p_u(\text{Kiribati}) \approx 0$. This tiny unigram probability for 'Kiribati' at first appears to do no harm, but voids the sole purpose of using unigram probabilities to counteract the higher sentence probability $p_m(s1)$ and $p_m(s2)$ in SLOR as we will show in the next section.

2.4 Sentence Probability

Now let us consider the sentence probability p_m from a 3-gram LM. Sentence probability is product of individual n-gram probabilities as described in Equation 2. Notice that all three sentences, s1, s2, and s3, in Table2 have a common prefix phrase '*He is a citizen of*' and differ only on the last word i.e., 'France', 'Tuvalu' and 'Kiribati'. The 3-gram LM will assign equal probabilities to all 3-grams ($p(a \mid He,\ is)$, $p(citizen \mid is,\ a)$, $p(of \mid a,\ citizen)$) within common prefix. Furthermore, based on our first assumption about words 'France' (i.e., high frequency) and 'Tuvalu' (i.e., rare) in the training corpus we should expect 3-gram probability $p(France \mid citizen,\ of)$ to be higher than $p(Tuvalu \mid citizen,\ of)$.

$$p_m\left(S\right) = p_m\left(w_1^n\right) = \prod_{t=1}^{n} p\left(w_t \mid w_{t-2}, w_{t-1}\right) \quad (2)$$

2.5 Incompetence of SLOR

In the ideal world, $p_m(s1)$ should be equal to $p_m(s2)$ since both are equally acceptable sentences. However, due to the above two observed outcomes, first, equal 3-gram probabilities for the common prefix on both s1 and s2; second, a higher 3-gram probability for the word 'France' will result in $p_m(s1)$ higher than $p_m(s2)$. This observation motivated Lau et al. (2017) to propose SLOR, where they counteracted this behavior by subtracting the unigram probabilities from sentence probabilities to get similar acceptability scores for equally acceptable sentences.

However, subtracting unigram probabilities does not solve problems for all different cases. Why? Recall in section 2.3, we discovered that for an OOV word 'Kiribati' unigram probability is tiny, i.e. p_u(Kiribati) ≈ 0. This tiny unigram probability for s3 will result in significantly different SLOR score for s3, therefore evaluating it as more acceptable sentence compared to s1 and s2. Hence $SLOR(s1) \approx SLOR(s2) \not\approx SLOR(s3)$ which is undesirable because the word choice ('France', 'Tuvalu' or 'Kiribati') should not lead to a different measure of acceptability.

3 Proposed Method

This observation that the lexical frequency of a word should not lead to a different measure of acceptability led us to think about how humans judge the acceptability of a sentence. We assert that human judgment of acceptability is only slightly influenced by word choice and is highly influenced by sentence structure.

Let us take an example from Table 3 to measure the acceptability of sentences s1 and s2. s1 and s2 are two original sentences with similar lengths and are equally acceptable. s3 and s4 represent sentences s1 and s2 with all identified NE spans and classified type as a subscript in prefix. s5 is the final transformed sentence (coarse-level representation) after replacing all identified NEs with the classified type for s1 and s2.

Index	Sentence
s1	Apple is set to hold its first event of the year on Tuesday.
s2	NEC is set to hold its second event of 2022 on Wednesday.
s3	[$_{\text{ORG}}$ Apple] is set to hold its [$_{\text{ORDINAL}}$ first] event of [$_{\text{DATE}}$ the year] on [$_{\text{DATE}}$ Tuesday].
s4	[$_{\text{ORG}}$ NEC] is set to hold its [$_{\text{ORDINAL}}$ second] event of [$_{\text{DATE}}$ 2022] on [$_{\text{DATE}}$ Wednesday].
s5	ORG is set to hold its ORDINAL event of DATE on DATE.

Table 3: Example sentences to explain the motivation and our proposed preprocessing data transformation strategy.

In Table 3 notice that sentences s1 and s2 are very similar in structure with few variations in the word choice, i.e. 'Apple' vs 'NEC', 'first' vs 'second', 'the year' vs '2022' and 'Tuesday' vs 'Wednesday'. Nonetheless, when it comes to a human judgment of acceptability for s1 and s2, one would rate both the sentences equally irrespective of different word type choices in a sentence. We argue that neither word choice nor lexical frequency should influence sentence acceptability. Therefore it does not matter if the word in the sentence is 'Apple' or 'NEC'; instead, the critical information is the fact that both the words refer to a single NE type, i.e., ORGANIZATION (ORG). Broadly we can think of any other ORG NE such as 'Microsoft', 'United Nations' etc, and it should not affect the measure of acceptability for the sentence. Similarly, the lexical frequency of phrases 'first' over 'second', 'the year' over '2022', and 'Tuesday' over 'Wednesday' is less critical than phrases referring to NE type ORDINAL, DATE, and DATE, respectively.

In a nutshell, to humans, the sentence's broad structure and transitions between POS (Lapata and Barzilay, 2005) are more critical than the lexical frequency of the words to determine the acceptability. Based on this motivation, if we were to replace phrases with their corresponding NE type, we can transform original (granular-level representation) sentences s1 and s2 to a standard (coarse-level representation) sentence s5. This transformation should help LMs overcome the issue of word choice and their lexical frequencies to influence sentences' measure of acceptability. If coarse-level representation is helpful, why not replace the complete sentence with the corresponding POS instead of only replacing proper nouns? The reason is that replacing proper nouns with their NE Type generates an advantageous representation. On the one hand, it abstracts away details that are not critical for determining acceptability; on the other hand, it retains original words and sentence structure that highly influence acceptability i.e., rest of the POS classes (verb, adjective, adverb, preposition, conjunction, and interjection). Now we propose our two-step (Step I and Step II) solution Replace Named Entities (RNE) for data transformation.

3.1 Step I: Named Entity Identification

First, we segmented a sentence into words using spacy's (Honnibal and Montani, 2017) NLP English word segmenter. After completing the segmentation process, we scan the segmented sentence sequentially to find the consecutive words that constitute a NE. We used spacy's statistical entity recognition system with a default trained pipeline to assign one out of eighteen types (e.g., companies, locations, organizations, and products.) to an identified NE.

3.2 Step II: Replacing words with Named Entity Types

After identifying both NEs and their respective types over segmented input sentences, we then start the replacement process. In this step, we replace one or more consecutive words in a sentence previously identified as a NE in step I with its corresponding identified NE type both for training and test corpus.

3.3 Complete Pipeline

After transforming all the sentences in the training and test corpus with RNE, we train n-gram LM over the transformed training corpus. Such a sentence transformation (both independent of LM and the acceptability metric) will provide a coarse-level representation of a sentence to help LM focus on the transitions of POS without worrying about the words chosen for NEs. Furthermore, we believe this will enable a LM to generalize better into new domains with different vocabulary. Moreover, this abstract representation of a sentence will help all probability-based metrics, including SLOR as shown in section 5.

Description	Size	Avg. Words	Avg. NE's	Avg. UNK's
BNC	5250	17.81	1.07	0.96
ENWIKI	2500	17.21	2.07	0.23
ADGER	300	7.30	0.53	0.04
ADGER-FILTERED	133	8.02	0.68	0.00

Table 4: Details of the test corpus. Description and Size represents name of dataset and total number of sentences in the dataset. Followed by average number of words, NEs, and UNK tokens per sentence respectively

4 Experiment Setup

4.1 Dataset

We adopt the BNC corpus (BNC Consortium, 2007) that comprises 6.07M sentences for training LM. Moreover, to show the effectiveness of our proposed data transformation strategy, i.e., RNE, we evaluated the trained LMs on sentences that exhibited varying degrees of acceptability. Based on previous work of (Lau et al., 2017), we evaluated LMs on four English language datasets (BNC, ENWIKI, ADGER, and ADGER-FILTERED) within the Statistical Model of Grammaticality (SMOG) (The Center for Linguistic Theory and Studies in Probability, 2015) test corpus. Table 4 shares the detailed statistics on test datasets. Each sentence in the test dataset is associated with a human judgement of acceptability for further details on collection of the human ratings refer to Appendix A.1.

4.2 Baselines

We first preprocessed the training corpus following Standard Preprocess (SP) protocol as described in (Lau et al., 2017). SP comprises of three steps, first is to segment the sentences, second, filter out sentences with fewer then threshold (seven) words, third, replace rare words (i.e. with frequency less than threshold of four) with an unknown (UNK) token.

4.3 Language Models

We trained three n-gram i.e., 2-gram, 3-gram, and 4-gram LMs on BNC corpus though preprocessed differently for baseline (only SP with a vocabulary of 104,950) and our proposed work (i.e., SP + RNE with a vocabulary of 100,688). Each LM (for both SP and SP + RNE) was trained with Kneser-Key (Kneser and Ney, 1995) smoothing method.

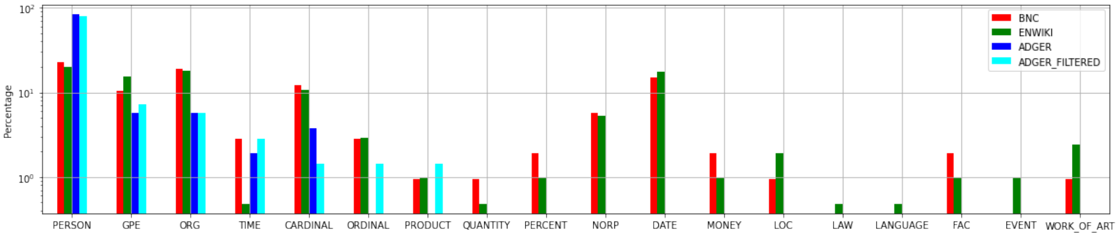


Figure 1: Percentage (Y axis in log scaled) of eighteen NE types (X axis) per sentence across four test datasets BNC, ENWIKI, ADGER and ADGER_FILTERED. Graph is sorted by the percentage of NE type on ADGER_FILTERED Dataset.

4.4 Metrics

To compare our results with previous work of Lau et al. (2017) we used pearson correlation between human judgement of acceptability and different probability scores (LogProb, Mean LP, Norm LP (DIV), Norm LP (SUB), SLOR) predicted to evaluate the performance of LMs. Due to the space limitation, we have only included the formula for SLOR in Section 2 for the formulation of rest of the metrics; refer to Appendix A.2.

Pearson Correlation We evaluated the performance of the LMs capability to predict the acceptability (X) by calculating it's pearson correlation with human judgement of acceptability (Y). In Equation 3 cov is the covariance. σ_X and σ_Y is the standard deviation of X and Y respectively.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

5 Results and Discussion

We now discuss the experimental results, findings and their implications on acceptability evaluation.

Performance Comparison: Table 5, 6, 7, and 8 shows the performance on BNC, ENWIKI, ADGER, and ADGER_FILTERED test datasets respectively. As the test datasets were already processed via SP, we only preprocessed test datasets with RNE to evaluate LMs trained via SP + RNE.

We observed that LMs trained via SP + RNE have a higher or equal correlation with human judgments on all the measures for BNC, ENWIKI, and ADGER test datasets compared to LMs trained only via SP. The only exception is 2-gram LM for the BNC test

Measure	**2-Gram**		**3-Gram**		**4-Gram**	
	SP	**SP+RNE**	**SP**	**SP + RNE**	**SP**	**SP + RNE**
LogProb	0.33	0.35	0.40	0.50	0.42	0.65
Mean LP	0.46	0.36	0.52	0.55	0.55	0.67
Norm LP (Div)	**0.53**	0.43	0.57	**0.62**	0.60	**0.73**
Norm LP (Sub)	0.23	0.13	0.29	0.30	0.33	0.44
SLOR	**0.53**	0.44	0.55	0.61	0.57	0.69

Table 5: Pearson's r of acceptability measure and mean sentence rating for BNC. For BNC all the metrics are multiplied by factor of 10.

Measure	**2-Gram**		**3-Gram**		**4-Gram**	
	SP	**SP+RNE**	**SP**	**SP + RNE**	**SP**	**SP + RNE**
LogProb	0.22	**0.28**	0.24	0.32	0.24	0.33
Mean LP	0.14	0.22	0.19	0.28	0.20	0.30
Norm LP (Div)	0.19	0.27	0.24	**0.33**	0.25	**0.35**
Norm LP (Sub)	0.01	0.01	0.07	0.07	0.08	0.10
SLOR	0.20	0.27	0.24	**0.33**	0.24	0.34

Table 6: Pearson's r of acceptability measure and mean sentence rating for ENWIKI

Measure	**2-Gram**		**3-Gram**		**4-Gram**	
	SP	**SP+RNE**	**SP**	**SP + RNE**	**SP**	**SP + RNE**
LogProb	0.06	0.07	0.07	0.08	0.08	0.08
Mean LP	0.07	0.07	0.09	0.09	0.09	0.10
Norm LP (Div)	0.11	0.11	0.13	0.13	0.13	**0.14**
Norm LP (Sub)	0.09	0.10	0.12	0.12	0.12	0.12
SLOR	**0.12**	**0.12**	**0.14**	**0.14**	**0.14**	**0.14**

Table 7: Pearson's r of acceptability measure and mean sentence rating for ADGER

Measure	2-Gram		3-Gram		4-Gram	
	SP	SP+RNE	SP	SP + RNE	SP	SP + RNE
LogProb	0.30	0.31	0.32	0.33	0.33	0.35
Mean LP	0.23	0.22	0.25	0.26	0.26	0.28
Norm LP (Div)	0.32	0.30	0.35	0.34	**0.36**	**0.36**
Norm LP (Sub)	0.16	0.10	0.20	0.15	0.23	0.18
SLOR	**0.34**	0.30	**0.36**	0.33	**0.36**	0.34

Table 8: Pearson's r of acceptability measure and mean sentence rating for ADGER FILTERED

dataset. Furthermore, we got mixed improvement results on the ADGER_FILTERED dataset.

Quantitatively we observed an improvement in the range of 3% (LogProb) to 52% (LogProb) for 2-gram and 4-gram LM, respectively. For ENWIKI, we observed an improvement in the range of 2% (Norm LP Sub) to 50% (Mean LP) for the 2-gram LM. For ADGER, we observed an improvement in the range of 1% (SLOR) to 13% (Mean LP) for 3-gram and 2-gram LM, respectively. For ADGER_FILTERED, we observed an improvement of 2% (Norm LP Div) to 11% (Log Prob) for 4-gram LM.

RNE's impact on probability metrics other than SLOR: We verified our hypothesis that neither word choice nor lexical frequency for NEs is critical in determining the acceptability of the sentence as we saw consistent improvement in correlation for all probability related measures in addition to SLOR.

Impact of NE count on correlation: We investigated the impact of NE count on the correlation improvement. All four test datasets exhibited different sentence characteristics. On the one hand, BNC and ENWIKI comprised 1.07 and 2.07 NEs per sentence; on the other hand, ADGER and ADGER_FILTERED only comprised 0.53 and 0.68 NEs per sentence. In other words, only half of the sentences in the test corpus have one NE. This observation indicates that the higher the number of NEs in the sentence bigger the improvement in correlation with human acceptability judgment. E.g., ENWIKI enjoyed the maximum number of NEs (2.07) per sentence, resulting in the maximum gain (50% for Mean LP) in correlation. The above result is aligned with our hypothesis since the higher the number of NEs in a sentence, the more abstract the sentence representation, resulting in less dependency on word choice and their lexical frequencies for acceptability evaluation.

Impact of NE Type on Performance: Fig. 1 shows the distribution of different NE types across four test datasets. BNC and ENWIKI displayed different sentence characteristics from ADGER and ADGER_FILTERED. Y-axis (log scaled) is the percentage of NE types over all NE count from the dataset for eighteen NE types (X-axis) across four test datasets. We observed that NE type PERSON was the most prominent, i.e., ≈ 80% for ADGER and ADGER_FILTERED vs. ≈ 20% for BNC and ENWIKI. Furthermore, not a single sentence in ADGER and ADGER_FILTERED possessed the following eleven NE types starting from QUANTITY, to EVENT, WORK_OF_ART as shown in Fig. 1 leading to different performance across four test datasets.

6 Qualitative Analysis

To give an intuition for our proposed methodology, we present one example, sentence s.157, from the ENWIKI test dataset in Table 9. With the full range of values, we apply a Z-score transformation to each of the values in Y (acceptability score) by subtracting the mean of Y from each of the values and dividing them by the standard deviation of Y. We applied the Z-score transformation on human acceptability ratings for the original sentence. Furthermore, for sentences preprocessed via SP and SP + RNE, we applied the Z-score to the SLOR scores predicted from 4-gram LM.

In the first sentence, there exist three NEs 'Myrtle Beach', 'Coast RTA', and 'Pee Dee Regional Transportation Authority'. SP + RNE replaces NE 'Myrtle Beach' with GPE, and 'Coast RTA', 'Pee Dee Regional Transportation Authority' with ORG. Consequently generating a coarse representation of the sentence, allowing LM to focus on the POS transition rather than being swamped by the long-phrase corresponding to NEs. Z-score of 1.025 from LM trained via SP + RNE is comparable to 1.033 for human ratings (with + sign signifying acceptable sentence), unlike the Z-score of -1.070 from LM trained on SP (with - sign signifying unacceptable sentence), which further supports our above claim.

Preprocessing	Sentence	Z-score
- (Original)	Myrtle Beach is served by the Coast RTA and the Pee Dee Regional Transportation Authority .	1.033
SP	myrtle beach is served by the coast UNK and the pee dee regional transportation authority .	-1.070
SP + RNE (Our's)	GPE is served by ORG and ORG .	1.025

Table 9: Sentence s.157 from ENWIKI dataset with unprocessed, standard and NE replacement preprocessing methods and it's corresponding Z-score. UP, SP and RNE corresponds to UnProcessed, Standard Preprossed, and our's proposed NE Replaced data preprocessing methodologies. UNK (i.e., Unknown) corresponds to word 'RTA' as OOV word.

7 Related Work

A series of recent successes of LMs on several NLP tasks have raised the critical question of automatic acceptability tests. Although, there have been several studies to access the acceptability automatically. However, there has not been enough effort to evaluate the impact of sentence representation's on different levels (i.e., granular vs. coarse) on acceptability.

Wan et al. (2005) was the first work to evaluate sentence acceptability independent of the source content. The authors suggested using grammatical judgments of a parser to assess the sentence acceptability. The motivation behind the idea was that if the parser is trained on the appropriate corpus, the poor performance of the parser on one sentence relative to the other sentence will suggest the presence of ungrammaticality and unacceptability. (Mutton et al., 2007) later extended the idea by training the machine learners on top of several parser outputs and showing its correlation with the human judgment of acceptability test. Unlike this line of work, we do not rely on the grammatical assessments from the parser but instead rely on the probabilities assigned by the LM for acceptability test.

Lau et al. (2017) proposed the first work to hold a probabilistic view on linguistic knowledge. They proposed and experimented on a comprehensive list of different probability-based metrics at the sentence level and word level. Taking this work forward (Kann et al., 2018) further introduced WPSLOR, a WordPiece-based version of SLOR, to reduce the LM size. Complementary to this work of exploring different probability-based metrics, we focused on varying levels of sentence representation (granular vs. coarse). Furthermore, we demonstrate that our data transformation strategy can lead to an additional gain in PCC measure between the metrics proposed by (Lau et al., 2017) and human acceptability judgments.

Motivated by centering theory (Grosz et al., 1995), Lapata and Barzilay (2005) argue that patterns of local entity transitions specify how the focus of discourse changes from sentence to sentence. To expose the entity transition patterns of readable texts, they represented a text by an entity grid and showed coherent texts exhibits certain regularities reflected in the topology of grid columns (i.e., discourse entities). This work inspires our idea to use NEs transition, but with few differences. First, our job is to evaluate the acceptability at the intra-sentence level, whereas their goal was to assess the coherence at the inter-sentence level. Second, they created a new representation of the input text in the form of the grid and trained ML learners on top of the grid to evaluate the coherence. However, we replaced the entities with their respective class in the sentence, used this coarse abstract representation, and relied on the LM to learn the acceptability measure from sentence structure.

Brown et al. (1992) presented a statistical algorithm for assigning words to classes (clusters) based on their frequency of co-occurrence with other words. Their method extracted classes with syntactic or semantic-based groupings of words and later proposed class-based n-gram LMs. Though our work is a specific instance of their approach, it answers two crucial questions for sentence transformation required for acceptability evaluation. First, which type of words to replace? Second, what should a word be replaced with? As we explained in section 2 only replacing proper nouns with NE Types generates required coarse-level representation without affecting the acceptability evaluation of the sentence. E.g., in the sentence "He sees the world from his *eyes*" randomly replacing the word 'eyes' with the other word 'mouth' from the same class or logical class name 'body parts' will drastically impact the acceptability measure of the sentence.

Pitler and Nenkova (2008) combined lexical, syntactic, and discourse features to produce a model to predict the human reader's judgments of text acceptability. They presented that discourse relations are strongly associated with the perceived quality of a text. Similarly, Vadlapudi and Katragadda (2010) proposed surface features like n-gram probabilities, 3-gram based class n-grams, hybrid model using both n-gram and class model on the POS-tag sequences and POS-chunk-tag sequences. They showed that their proposed models, especially the hybrid approach on the POS-chunk-tag sequence, can highly correlate with the human judgment of acceptability. Unlike this line of work, we kept the focus on NEs and proposed a data transformation methodology independent of both the LM and the metric used to measure the correlation with human acceptability judgment.

8 Conclusion

Several probability metrics have been proposed to conduct the reference less acceptability evaluation of a sentence automatically. SLOR in particular, has gained popularity given it removes the impact from the confounding factors of noise like sentence length and lexical frequency. We assert that the issues related to word choice and its lexical frequency persist for SLOR. We proposed a data preprocessing strategy motivated by humans who evaluate the sentence based on sentence structure and transitions between POS at a coarser sentence representation. RNE, our proposed method, identifies all NEs in a sentence and replaces it with the classified type of NE. Based on the results of the experiments, we found a correlation between NEs count in a sentence and improvement in LMs acceptability score. We observed an improvement (up to 52%) or equal performance for three (i.e., BNC, ENWIKI, and ADGER) out of four English datasets. In this work, we only focused on one class of POS, i.e., NE. In the future, we would like to find an optimal dynamic representation of a sentence based on its content to help LM predict acceptability scores better.

References

David Adger. 2003. Core syntax: A minimalist approach.

BNC Consortium. 2007. The british national corpus, version 3 (bnc xml edition). distributed by oxford university computing services on behlaf of the bnc consortium. `http://www.natcorp.ox.ac.uk/`, Last accessed on 2022-03-21.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Comput. Linguistics*, 18:467–479.

Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Comput. Linguistics*, 21:203–225.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *CoNLL*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1:181–184 vol.1.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41 5:1202–1241.

David Manning. 2002. Probabilistic syntax.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic, June. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*.

Neomy Storch. 2009. The impact of studying in a second language (l2) medium university on the development of l2 writing. *Journal of Second Language Writing*, 18:103–118.

The Center for Linguistic Theory and Studies in Probability. 2015. Statistical model of grammaticality. distributed by the center for linguistic theory and studies in probability. `https://gu-clasp.github.io/projects/smog/experiments/`, Last accessed on 2022-03-21.

Ravikiran Vadlapudi and Rahul Katragadda. 2010. On automated evaluation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 7–12, Los Angeles, CA, June. Association for Computational Linguistics.

Stephen Wan, R. Dale, and Mark Dras. 2005. Searching for grammaticality: Propagating dependencies in the viterbi algorithm. In *ENLG*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: A benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A Metrics

A.1 Test Dataset Details

Humans annotated SMOG on three modes of presentation. (a) binary (MOP2), where annotators choose between two options: unacceptable and acceptable (b) four-category (MOP4), where annotators choose between four options: highly unacceptable, somewhat unacceptable, somewhat acceptable, and highly acceptable. (c) a sliding scale (MOP100) with two extremes, highly unacceptable and highly acceptable.

BNC test corpus comprised of 500 random sentences (in English) from the BNC training corpus with a length of 8-25 words. These 500 sentences were machine-translated (using Google Translate) to four target languages, i.e., Norwegian, Spanish, Chinese, and Japanese, and then back to English. This led to 2500 sentences, i.e. 500 *en* original and 500 each from back translation of original *en* to four target languages i.e, *es, no, zh, ja* and then back to en. BNC test corpus comprised 5250 annotated sentences, 2500 on MOP2, 2500 on MOP4, and 250 (10% randomly selected from 2500) on MOP100. To keep the training and test corpus distinct, we removed 500 English test sentences from the BNC training corpus.

Furthermore, SMOG comprised of linguistic sentences adopted from (Adger, 2003)'s syntax textbook. Lau et al. (2017) selected 100 random sentences from (Adger, 2003) where half of them were good (grammatical on author's judgment), and half of them starred (ungrammatical on author's judgment). To focus on syntactic violations, authors created another dataset, ADGER-FILTERED, after filtering out all sentences from (Adger, 2003) that were semantically or pragmatically anomalous. So that the left sentences only consisted of sentences that are either syntactic well-formed or syntactic violations.

A.2 Sentence Level Metrics Formulations

Log Probability In Equation 4 LogProb relates to the log of the sentence probability assigned by the LM.

$$\text{LogProb} = \log p_m(S) \tag{4}$$

Mean Log Probability In Equation 5 Mean LP relates to mean (i.e. average) log of the sentence probability. Which is calculated by dividing the log of the sentence probability with the length of the sentence.

$$\text{Mean LP } = \frac{\log p_m(S)}{|S|} \tag{5}$$

Normalized Log Probability Division In Equation 6 Norm LP (Div) relates to the normalized log probability which is calculated by diving the log of sentence probability with log of the sentence unigram probability.

$$\text{Norm LP (Div) } = -\frac{\log p_m(S)}{\log p_u(S)} \tag{6}$$

Normalized Log Probability Subtraction In Equation 7 Norm LP (Sub) relates to the normalized log probability which is calculated by subtracting the log of the sentence probability with log of sentence unigram probability. Which is also same as log of the division of the sentence probability with the sentence unigram probability.

$$\begin{aligned}\text{Norm LP (Sub)} &= \log p_m(S) - \log p_u(S) \\ &= \log \frac{p_m(S)}{p_u(S)}\end{aligned} \tag{7}$$

Text Classification Using a Graph Based on Relationships Between Documents

Hiromu Nakajima
Graduate School of Sci. and Eng., Ibaraki Univ./ 4-12-1 Nakanarisawacho, Hitachi City, Ibaraki Prefecture, 316-8511, Japan
22nm738g@vc.ibaraki.ac.jp

Minoru Sasaki
Graduate School of Sci. and Eng., Ibaraki Univ./ 4-12-1 Nakanarisawacho, Hitachi City, Ibaraki Prefecture, 316-8511, Japan
minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract

Text classification, which determines the genre of a document based on cues such as the co-occurrence of words and their frequency of occurrence, has been studied in various approaches to date. Conventional text classification methods using graph-structured data express relationships between words and between words and documents in the form of weights of edges between each node. Then, the graph is input to a graph neural network for learning. However, conventional methods do not represent the relationship between documents on the graph, and thus cannot directly consider the relationship between documents. Therefore, we propose a text classification method using the graph considers the relationships among documents. This method directly expresses the relationship between documents by adding the similarity of documents as weights of edges between document nodes to the graph of the conventional method. The constructed graph is then input to a graph convolutional neural network for learning. We conducted experiments using five English corpus (20NG, R52, R8, Ohsumed, and MR) to evaluate proposed method. The results show that the proposed method improves accuracy compared to the conventional method and that the use of relationships among document nodes is effective. Experimental results also show that the proposed method is particularly effective on datasets with relatively long documents.

1 Introduction

Text classification is the task of estimating an appropriate label for a given document from a set of predefined labels. This task is one of the fundamental problems in natural language processing. This technique has been applied in the real world to automate the task of document classification by humans. Many researchers are interested in developing applications that leverage text classification methods such as Junk mail classification, topic labeling and sentiment analysis.

In the past few years, convolutional neural networks that can take advantage of graph structural information have been used in solving text classification problems. TextGCN (Yao et al., 2019) is one of the examples of graph-based text classification methods. In TextGCN, word nodes and document nodes are represented on the same graph, which is input to GCN for learning. VGCN-BERT (Lu et al., 2020) trained by constructing a graph based on word co-occurrence information and word embedding representation of BERT and inputting the graph to GCN. RoBERTaGCN (Yuxiao et al., 2021) is a text classification method that combines the benefits of GCN's transductive learning with the knowledge gained from BERT's large-scale prior learning using large amounts of unlabeled data. This method boasts the best performance among existing methods for text classification with four datasets: 20NG, R8, Ohsumed, and MR. The graphs in these text classification methods use word-to-word and word-to-document relationships. However, conventional graph-based text classification methods do not use the relationship between documents. Therefore, we thought that accuracy could be improved by using the relationship between documents.

In this study, we aimed to solve the problem of Conventional graph-based text classification methods described above paragraph by adding relations between documents to the edges between document nodes, and to improve the classification performance of RoBERTaGCN. Specifically, we input each document into the BERT model and obtain the vector of '[CLS]' token in final hidden layer. Then, we calculated the cosine similarity of these '[CLS]' token vectors of each document and added the cosine similarity that exceeded a predetermined threshold as a weight between document nodes. Then, we can create an effective graph structure that considers the relation between document nodes to improve the accuracy in each dataset of RoBERTaGCN. In addition, we consider that topic drift is less likely to occur because document information can be propagated without going through word nodes.

2 Related Work

Graph neural networks (Scarselli et al., 2008) are neural networks that learn relationships between graph nodes via the edges that connect them. There are several types of GNNs. The graph convolutional networks (Kipf and Welling, 2016a) takes a graph as input and learns the relationship between the nodes of interest and their neighbors through convolutional computation using weights assigned to the edges between the nodes. The graph autoencoder (Kipf and Welling., 2016b) is the autoencoder that extracts important features by dimensionally collapsing the input data. In Graph Attention Network (Velickovi et al., 2017), the weights of edges between nodes and the coefficients representing the importance of neighboring nodes are used to extract important features. GNNs have been used in a wide range of tasks in the field of machine learning, such as relation extraction, text generation, machine translation and question answering. In the field of machine learning, GNNs have been used in a wide range of tasks and have demonstrated high performance. The success of GNNs in these wide range of tasks has motivated us to study text classification methods using GNNs. In TextGCN (Yao et al., 2019), document nodes and word nodes are represented on the same graph (heterogeneous graph), which is input to the GCN for training. Recently, there has been a lot of research on text classification methods that combine large scale pre-training models such as BERT with GNNs. VGCN-BERT (Lu et al., 2020) trained by constructing a graph based on word co-occurrence information and word embedding representation of BERT and inputting the graph to GCN. In BertGCN (Yuxiao et al., 2021), the heterogeneous graph of words and documents is constructed based on word co-occurrence information and BERT's document embedding representation, and the graph is input to GCN for learning. A detailed description of BertGCN is given in the next chapter.

3 RoBERTaGCN

BertGCN is a text classification method that combines the transductive learning of GCN with the knowledge gained from large-scale pre-training using large amounts of unlabeled data in BERT. BertGCN is trained by inputting each document into BERT, extracting document vectors from its output, and inputting them into GCN as initial representations of document nodes along with heterogeneous graphs of documents and words. BertGCN has now achieved state-of-the-art in the text classification task.

In BertGCN, the weights between nodes on a heterogeneous graph of words and documents are defined as shown in Equation (1) below. PPMI is used as the weights between word nodes, and TF-IDF is used as the weights between word and document nodes. As shown in equation (1), BertGCN does not express the relations between document nodes as the form of edge weights between nodes.

$$A_{i,j} = \begin{cases} PPMI(i,j), & i,j \; are \; words \; and \; i \neq j \\ TF - IDF(i,j), & i \; is \; document, j \; is \; word \\ 1, & i = j \\ 0, & otherwise \end{cases} \quad (1)$$

Yuxiao Lin et al. distinguish the names of the training models depending on the type of GNN and the pre-trained models of BERT. The names of the models are listed in Table 1. In this study, we targeted RoBERTaGCN for improvement.

Pre-Trained Model	GNN	Name of Model
bert-base	GCN	BertGCN
roberta-base	GCN	RoBERTaGCN
bert-base	GAT	BertGAT
roberta-base	GAT	RoBERTaGAT

Table 1. Names of the Models

4 Method

First, we construct a heterogeneous graph of words and documents using each document. Next, we input the graph information (weight matrix and initial node feature matrix) to BERT and GCN and obtain each prediction. Finally, we calculate the linear interpolation of each prediction and adopt the result as the final prediction.

4.1 Build Heterogeneous Graph

First, we build a heterogeneous graph containing word nodes and document nodes. The weights of the edges between nodes i and j are defined as in Equation (2).

$$A_{i,j} = \begin{cases} COS\_SIM(i,j), & i,j \text{ are documents and } i \neq j \\ PPMI(i,j), & i,j \text{ are words and } i \neq j \\ TF-IDF(i,j), & i \text{ is document}, j \text{ is word} \\ 1, & i = j \\ 0, & otherwise \end{cases} \tag{2}$$

In RoBERTaGCN, as shown in Equation (1), the relation between words and the relation between words and documents were considered as the form of edge weights between nodes, but the relation between documents was not considered. Therefore, we improved RoBERTaGCN to consider the relation between documents by expressing the relation between documents as the form of edge weights between document nodes. $COS\_SIM(i,j)$ in equation (2) is the weight of the edge between document nodes and represents the cosine similarity. Specifically, we added the weights of the edges between document nodes by following the steps I to III below.

I. tokenize each document

Each document is tokenized by the BertTokenizer and converted into a sequence of tokens that can be input to BERT. If the number of words in a document exceeds the BERT input limit of 512, including special tokens, 510 words were extracted from the front of the document and used.

II. obtain the CLS vector

Each tokenized document is entered into BERT to obtain the CLS vector at its final hidden layer, which is a vector reflecting the features of the entire document.

III. calculate and add cosine similarity

Calculate the cosine similarity between the CLS vectors of each acquired document. If the obtained cosine similarity exceeds a predetermined threshold, the cosine similarity is added as the weight of the edge between the corresponding document nodes.

We used positive mutual information (PPMI) for weight of the edges between word nodes. We used TF-IDF for weight of the edges between word nodes and document nodes. The process from the second section onward is in accordance with RoBERTaGCN.

4.2 Creating the Initial Node Feature Matrix

The next step is to create the initial node feature matrix to be input to the GCN. We use BERT to obtain document embedding representations and treat them as input representations of document nodes. The embedding representation X_{doc} of a document node is represented by $X_{doc} \in \mathbb{R}^{n_{doc} \times d}$ using the number of documents n_{doc} and the number of embedding dimensions d. In general, the initial node feature matrix is given by the following equation (3).

$$X = \begin{pmatrix} X_{doc} \\ 0 \end{pmatrix}_{(n_{doc}+n_{word}) \times d} \tag{3}$$

4.3 Input to GCN and Learning by GCN

The weights of the edges between nodes and the initial node feature matrix shown in equations (2) and (3) are input to the GCN for training. The output feature matrix $L^{(i)}$ of the i-th layer is calculated by Equation (4).

$$L^{(i)} = \rho\left(\tilde{A} L^{(i-1)} W^{(i)}\right) \tag{4}$$

ρ is the activation function, $\tilde{A}$ is the normalized adjacency matrix. $W^{i} \in \mathbb{R}^{d_{i-1} \times d_i}$ is the weight matrix at layer i, $L^{(0)}$ is X, which is the input feature matrix of the model. The output of the GCN is treated as the final representation of the document nodes, and its output is input to the softmax function for classification. The prediction by the output of GCN is given by equation (5). g represents the GCN model.

$$Z_{GCN} = softmax\big(g(X, A)\big) \tag{5}$$

4.4 Interpolation of Predictions with BERT and GCN

We optimize the GCN with an auxiliary classifier that directly handles the BERT embedded

Dataset	Number of Documents	Average of Words	Training Data	Test Data
20NG	18846	206.4	11314	7532
R8	7674	65.7	5485	2189
R52	9100	69.8	6532	2568
Ohsumed	7400	129.1	3357	4043
MR	10662	20.3	7108	3554

Table2. Information of Each Data Set

representation for faster convergence and better performance. Specifically, we create an auxiliary classifier with BERT by feeding the document embedding representation X and the weight matrix W directly into the softmax function. The prediction by the auxiliary classifier is given by the following equation (6).

$$Z_{BERT} = softmax(WX) \tag{6}$$

Then, a linear interpolation is performed using Z_{GCN} which prediction from RoBERTaGCN and Z_{BERT} which prediction from BERT, and the result of the linear interpolation is adopted as the final prediction. The result of linear interpolation is given by equation (7).

$$Z = \lambda Z_{GCN} + (1 - \lambda) Z_{BERT} \tag{7}$$

λ controls the trade-off between the two predictions, meaning that if $\lambda = 1$, we use the full RoBERTaGCN model, and if $\lambda = 0$, we use only the BERT module. $\lambda \in (0, 1)$, we can balance the predictions from both models and RoBERTaGCN model can be more optimized. $\lambda = 0.7$ is the optimal value of λ, as shown by the experiments of Yuxiao.

5 Experiments

We evaluated the classification performance of the proposed method by conducting experiments with the cosine similarity threshold set between 0.5 and 0.95 to 0.995 in increments of 0.005 and investigated the optimal cosine similarity threshold for each data set.

5.1 Dataset

We evaluated the performance of the proposed method by conducting experiments using the five data sets shown in Table 2. We used the same data used in RoBERTaGCN. Each dataset was already divided into training and test data, which we used as is.[1] The number of data for training and test data is shown in Table 2.

・20-Newsgroups(20NG)

20NG is a dataset in which each document is categorized into 20 news categories, and the total number of documents is 18846. In our experiments, we used 11314 documents as training data and 7532 documents as test data.

・R8, R52

Both R8 and R52 are subsets of the dataset provided by Reuters (total number is 21578). R8 has 8 categories and R52 has 52 categories. The total number of documents in R8 is 7674, and we used 5485 documents as training data and 2189 documents as test data. The total number of documents in R52 is 9100, and we used 6532 documents as training data and 2568 documents as test data.

・Ohsumed

This is a dataset of medical literature provided by the U.S. National Library of Medicine, and total number of documents is 13929. Every document has one or more than two related disease categories from among the 23 disease categories. In the experiment, we used documents that had only one relevant disease category, and the number of documents is 7400. We used 3357 documents as training data and 4043 documents as test data.

・Movie Review(MR)

This is a dataset of movie reviews and is used for sentiment classification (negative-positive classification). The total number of documents was 10662. We used 7108 documents as training data and 3554 documents as test data.

[1] https://github.com/ZeroRin/BertGCN/tree/main/data

GPU	Tesla V100（SXM2） / A100（SXM2）
Memory	12.69GB（standard） / 51.01GB（CPU / GPU(high memory)） / 35.25GB（TPU(high memory)）
Disk	225.89GB（CPU / TPU） / 166.83GB（GPU）

Table3. Details of the Specifications of Google Colaboratory Pro+

	20NG	R8	R52	Ohsumed	MR
Text GCN	86.34	97.07	93.56	68.36	76.74
Simplified GCN	88.50	-	-	68.50	-
LEAM	81.91	93.31	91.84	58.58	76.95
SWEM	85.16	95.32	92.94	63.12	76.65
TF-IDF+LR	83.19	93.74	86.95	54.66	74.59
LSTM	65.71	93.68	85.54	41.13	75.06
fastText	79.38	96.13	92.81	57.70	75.14
RoBERTaGCN	89.15	98.58	94.08	72.94	88.66
0.5	×	49.47	×	64.73	×
0.95	89.29	98.26	92.83	73.73	88.21
0.955	89.42	98.63	94.08	72.74	88.21
0.96	89.74	98.49	**94.16**	73.49	88.52
0.965	89.54	98.45	93.15	**74.13**	88.15
0.97	89.43	98.45	93.77	73.41	88.66
0.975	**89.82**	98.63	93.57	73.49	**89.00**
0.98	89.60	98.54	93.96		88.29
0.985	89.64	98.54	92.95	73.46	88.58
0.99	89.76	98.36	93.42	73.71	88.55
0.995	89.51	**98.81**	93.26		88.31

Table4. Result of Experiment

5.2 Experimental Environment

The experiments were conducted using Google Colaboratory Pro+, an execution environment for Python and other programming languages provided by Google. The details of the specifications of Google Colaboratory Pro+ are shown in Table 3.

We experimented by setting the threshold of cosine similarity between 0.5 and 0.95 to 0.995 in increments of 0.005 when adding the cosine similarity of CLS vectors as the weight of edges between document nodes. The performance of the proposed method was evaluated by verifying the prediction results with test data and obtaining the percentage of correct answers.

5.3 Result of Experiment

The result of experiment for each threshold of cosine similarity are shown in Table 4, along with the correct response rate of the original RoBERTaGCN.

The items marked as × are experiments could not be completed due to lack of memory. Items marked with "-" are those for which the percentage of correct responses was not indicated in the original paper. In experiment with Ohsumed, the experiments with threshold of 0.99 and 0.995, and threshold of 0.975 and 0.98 had the same number of edges of the cosine similarity of CLS vectors, so they are denoted together. It was confirmed that the proposed method outperformed the original RoBERTaGCN on all datasets at certain thresholds, but for R8, R52, and MR, there were only one or two thresholds where the proposed method outperformed the original RoBERTaGCN. On the other hand, 20NG outperformed the original RoBERTaGCN at all thresholds from 0.95 to 0.995, and Ohsumed also outperformed the original RoBERTaGCN at most of the thresholds. Most notably, the experiment with 20NG of threshold 0.975 outperformed the original RoBERTaGCN by 0.67% and the experiment with Ohsumed of

Dataset	pmi Edge	tf-idf Edge	cos_sim Edge	Average of Cosine Similarity
20NG	22413246	2276720	×	0.838
R8	2841760	323670	29441186	0.846
R52	3574162	407084	41400215	0.840
Ohsumed	6867490	588958	27376155	0.837
MR	1504598	196826	56674250	0.823

Table5. Number of Various Edges Added and the Average of Cosine Similarity

Dataset	Total Number of Document Node Combinations	Number of cos_sim Edges Added	Percentage of Edges Added
20NG	177576435	753	0.0004240
R8	29441301	175	0.0005944
R52	41400450	28890	0.0697818
Ohsumed	27376300	15	0.0000547
MR	56833791	921	0.0016205

Table 6. Percentage of the Number of Edges Added

threshold 0.965 outperformed the original RoBERTaGCN by 1.19%.

6 Discussion

Table 5 shows the number of various edges added and the average of cosine similarity in each data set. The item marked as × is the experiment could not be completed adding weight due to lack of memory. In the experiment where the threshold was set to 0.5, the experiment could not be completed due to lack of memory in the datasets of 20NG, R52, and MR. Even for R8 and Ohsumed, which were able to complete the experiment, the classification performance was much lower than that of the original RoBERTaGCN. The reason for both is that the number of edge weights between document nodes to be added became too large. In all datasets, the number of edges of cosine similarity is more than twice as large as the number of PMI edges and TF-IDF edges. In addition, since the average of the cosine similarity of the CLS vector is between 0.8~0.85 in all datasets, it is thought that a huge number of weights of edges between document nodes that are not in the same genre are also added, and they have become noise.

Analyzing the average number of words for each dataset in Table 2 and the experimental results in Table 4, we can see that the proposed method tends to obtain higher classification performance for datasets with higher average number of words compared to the original RoBERTaGCN. We believe this is because the higher the average word count, the better the CLS vectors of the documents reflect the features of those documents and the more cosine similarity weights are added between document nodes of the same genre. On the other hand, the lower the average number of words, the less the difference in the CLS vectors of the documents, the higher the cosine similarity of the CLS vectors of the documents in different genres, and the more cosine similarity weights were added to the weights between the nodes of the documents in different genres. This is thought to be the reason why the classification performance did not improve as expected in experiments with dataset have lower the average number of words.

We calculated the percentage of the number of added cosine similarities at the threshold of the cosine similarity of the CLS vector that shows the highest classification performance in Table 4. The calculation results are shown in Table 6.

Since there is no relationship between the percentage of the number of added edges and the classification performance, we think it is necessary to conduct future experiments using criteria such as "upper XX% of the cosine similarity value", instead of using the threshold of the cosine similarity of the CLS vector to determine the weights to be added between document nodes, to clarify the relationship between the number of edges between document nodes and the classification performance.

7 Conclusion and Future Work

In this paper, we confirmed that RoBERTaGCN can be improved by adding the cosine similarity of CLS vectors of documents as weights of edges between document nodes, and that it outperforms the classification performance of the original

RoBERTaGCN. In particular, experiments show that the proposed method is effective for long documents.

In the future, we intend to study the compatibility of the proposed method with GAT and the optimal value of the parameter λ for linear interpolation.

References

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. IEEE Transactions on Neural Networks, 20(1):61–80.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima's an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1957-1967, Copenhagen, Denmark. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 7370-7377.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. arXiv preprint arXiv:1910.02356.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pages 452-455. IEEE.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao. 2021. Deep Learning Based Text Classification: A Comprehensive Revie. arXiv:2004.03705v3

Thomas N Kipf and Max Welling. 2016b. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308.

Thomas N Kipf and Max Welling. 2016a. Semisupervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor graph convolutional networks for text classification

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li and Fei Wu. 2021. BertGCN: Transductive Text Classification by Combining GCN and BERT.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In European Conference on Information Retrieval, pages 369-382. Springer.

Speaker Identification of Quotes in Japanese Novels based on Gender Classification Model by BERT

Yuki Zenimoto, Takehito Utsuro
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba,
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
s2220753_@_s.tsukuba.ac.jp, utsuro_@_.tsukuba.ac.jp

Abstract

In the Japanese language, particles and auxiliary verbs in utterances tend to vary depending on the speaker's gender. Thus, the linguistic expressions within the utterance are an important hint in identifying the speaker. This research proposes a method for identifying the speaker in novels using linguistic expressions within the utterances that reflect the gender. We constructed a dataset of utterances with gender-specific linguistic expressions by automatically collecting utterances from novels that contained the first-person pronouns "俺 (ore)" or "私 (watashi)", considering that "俺 (ore)" is primarily used by males and "私 (watashi)" is primarily used by females. We fine-tuned a BERT (Devlin et al., 2019)-based gender-specific language model that classifies the gender of the speaker of a given utterance using this dataset as the training data. The fine-tuned gender classification model achieved an accuracy of 87.9% when evaluating the utterances of two main characters in a Japanese romance novel, demonstrating that this gender classification model is effective in speaker identification.

1 Introduction

Novels are valuable resources for enriching dialogue systems. Various characters converse with each other in novels, and their behaviors are described in the narratives. If such information is appropriately extracted, it can be used in dialogue systems and various research such as personality analysis. Dialogue systems are required to respond appropriately to the users' utterances, and it is also expected to have a specific personality. To train dialogue models, it is necessary to collect dialogue data from web services such as Twitter and Reddit (Serban et al., 2015; Mazaré et al., 2018). However, these dialogue data are insufficient for training dialogue models capable of acting as if they have specific personalities. This problem is simply because these dialogue data consist of a mixture of utterances of numerous personalities and are difficult to utilize as a source for building dialogue models with specific personalities. Considering the preceding discussion, it is required to prepare dialogue data that are accompanied by profiles (Zhang et al., 2018; Sugiyama et al., 2021). Crowdsourcing is commonly used to collect such dialogue data with profile information. However, crowdsourcing large scale dialogue data with profile information is extremely expensive.

Considering this limitation, this paper introduced an approach for gathering dialogue data from novels, examining the task of speaker identification in novels to generate persona data. Previous work on speaker identification in English novels (O'Keefe et al., 2012; Muzny et al., 2017) relied heavily on narrative elements surrounding the target utterance and the utterances before and after it. However, the target utterance itself was not effectively used. Here, it is important to note that, in the Japanese language, particles and auxiliary verbs in utterances differ depending on the speaker's gender (Miyazaki et al., 2015; Murai, 2018). Thus, the linguistic expressions within the utterance are also an essential clue in speaker identification.

This paper proposes a method to identify the

First-Person Pronouns	Example Sentences	Significant Words
“俺” (ore) (typical for males)	俺の番だな ore no ban da **na** (It is my turn)	na, ze
	俺は家に戻るぜ ore wa ie ni modoru **ze** (I am going home)	
“私” (watashi) (typical for females)	私の番だね watashi no ban da **ne** (It is my turn)	ne, wa
	私は家に戻るわ watashi wa ie ni modoru **wa** (I am going home)	

Table 1: Significant Words for First-Person Pronouns “俺 (ore) ” and “私 (watashi)”

speaker in novels, which uses the linguistic expressions within the utterance that reflect the gender. We gathered utterances within novels containing gender-specific linguistic expressions on a large scale, concentrating on the first-person pronouns within sentences as signals. In the Japanese language, there are numerous varieties of personal pronouns. Personal pronouns in Japanese can approximately indicate various attributes, such as gender, age, temperament, and characters’ social status. Significantly, the first-person pronouns “俺 (ore)” and “私 (watashi)” are remarkably contrasting, with “俺 (ore)” being primarily used by males and “私 (watashi)” being primarily used by females. As shown in Table 1, terms commonly used by males frequently occur in utterances containing “ 俺 (ore)”, whereas terms commonly used by females frequently appear in utterances containing “私 (watashi)”. Thus, we constructed a dataset of utterances containing gender-specific language expressions by automatically collecting utterances from novels that contained the first-person pronouns “俺 (ore)” or “私 (watashi)”. We fine-tuned a BERT (Devlin et al., 2019)-based gender-specific language model that classifies the speaker’s gender of a given utterance using linguistic expressions within the utterance as clues[1]. In this research, we refer to this gender-specific language model as the *gender classification model.*

The fine-tuned gender classification model was evaluated against the utterances of two main characters in a Japanese romantic novel. The results showed that the fine-tuned gender classification model achieved an accuracy of 87.9%, indicating that this model is effective in speaker identification. Additionally, we found that the speaker alternation constraint employed in previous studies (He et al., 2013) and the constraint on personal nouns within utterances improve the speaker identification performance.

2 Related Works

Previous studies on constructing Japanese persona datasets have usually relied on crowdsourcing to gather sentences that reflect a given personality. Sugiyama et al. (2021) used crowdsourcing to create 100 different personalities and collected a total of 61,794 utterances in 5,000 dialogues. They set the unit price at 300 yen (approximately three dollars) per dialogue. Ishii et al. (2021) proposed the use of role play-based question-answering to efficiently gather paired utterances expressing an anime character’s personality. Users who were familiar with the anime character provided them with 15,112 of these utterance pairs for free. However, finding relevant users who are familiar with such characters and then collecting utterance pairs using the same cost-free method are difficult.

Two corpora that could be used for speaker identification in Japanese novels are Aozora Bunko[2] and the balanced corpus of contemporary written Japanese (BCCWJ) (Maekawa et al., 2014). Aozora Bunko is a Japanese digital library that contains thousands of out-of-copyright Japanese works. BCCWJ[3] is a Japanese first 100 million words balanced corpus covering 2,663 novels (published between 1976 and 2005) and 270,388 individual utterances,

[1] In both training and testing, the first-person pronouns “俺 (ore)” and “私 (watashi)” of the input sentence were replaced with the [MASK] token.

[2] https://www.aozora.gr.jp/

[3] https://clrd.ninjal.ac.jp/bccwj/

Example Sentences	Speaker (gender)	Personal Nouns (Japanese/Pronunciation/English)
わたしは背が高すぎて、 お姉様の服は着られませんから...... (I am too tall to wear your clothes. ......)	Marie (female)	わたし/“watashi”/I お姉様/“one-sama”/sister
......キュロス様は、今、どちらに......？ (Where ...... is Mr. Kyuros now ......?)		キュロス様/“Kyurosu-sama”/Mr. Kyuros
どうもありがとう、使用人のお嬢さん。 (Thank you very much, servant girl.)	Kyuros (male)	お嬢さん/“ojo-san”/girl
おはよう、マリー。 (Good morning, Marie.)		マリー/“mari”/Marie

Table 2: Examples of the Annotation of Utterances Within the Novel for Evaluation

each linked to the speaker and some attributes (gender, age group, occupation, and so on) (Yamazaki et al., 2018). Both Aozora Bunko and BCCWJ offer collections of older novels, and many utterances are written in a different style compared with those in contemporary novels. Therefore, these corpora are not appropriate for gathering distinctive characters' utterances, such as those in anime and comics.

Early works on quote attribution in English novels focused on textual indications to determine the mention corresponding to the speaker of a quote (Elson and McKeown, 2010; Muzny et al., 2017). They mainly used patterns like quote-mention-verb and dependency parses to extract mentions and speakers. Only vocatives were extracted from the utterance and were incorporated in the classification of the next speaker or the listener as features within the utterance (He et al., 2013; Yeung and Lee, 2017); however, linguistic styles such as auxiliary verbs were not used in the task for English novels.

In the field of a character's linguistic style analysis, Miyazaki et al. (2016) identified 13 categories of linguistic peculiarities that can be used to identify the linguistic styles of most Japanese fictional characters. Akama et al. (2018) proposed style-sensitive word vectors that capture the stylistic similarity between two words. To measure the intensity of the persona characteristics of an utterance, Miyazaki et al. (2021) proposed a persona speaker probability that distinguishes the persona of the speaker of each utterance. In contrast to their method that classifies utterances into specific characters, our proposed method classifies the speaker of the utterances as male or female.

3 The Novel for Analysis

To evaluate the performance of speaker identification by the gender classification model and other additional constraints, we selected a contemporary novel[4] with a romance theme between two main characters — an aristocratic man (“Kyuros”) and an aristocratic woman (“Marie”).

3.1 Annotation

The first author of the paper is the annotator of our dataset. Quotes in Japanese text were represented by “「” at the start and “」” at the end, and we describe such utterance as a quote. Table 2 shows examples of the annotation. The annotator was instructed to associate the following attributions to each utterance.

- **Speaker Name of the Utterance**
  The speaker of the utterance was identified by a unique name. Utterances by characters whose names were not mentioned in the novel, such as store clerks and crowds, were labeled as “Others”.

- **Personal Nouns appearing within the Utterance**
  Personal nouns, such as first- and second-person pronouns, were extracted from the utterance. Personal nouns such as “会社員 (kaishain)” (company employee) and “使用人 (shiyonin)” (servant) were not tagged in this context because they were mentioned in the same way by any speaker. However,

[4] https://ncode.syosetu.com/n1860fv/

Type		Number of dialogues	Number of utterances by Marie (female) and Kyuros (male)	Number of utterances by characters other than Marie (female) and Kyuros (male)
Dialogues consisting of utterances only by Marie (female) and Kyuros (male)	Two speakers take turns alternately (speaker alternation constraint *satisfied*)	298	900	0
	One speaker continues more than one utterance (speaker alternation constraint *unsatisfied*)	(1) (consisting of utterances by one speaker only)	2	0
Total		298 (+1)	902	0
Dialogues including characters other than Marie (female) or Kyuros (male)		504	821	851
Isolated single utterance immediately preceded/followed by narratives		0	485	426
Total		504	1,306	1,277

Table 3: Statistics of the Dataset [Focusing on utterances by "Marie" (female) or "Kyuros" (male), where we exclusively evaluate utterances by "Marie" (female) or "Kyuros" (male) in this research. A total of 902 utterances of the upper half of the table were evaluated with the speaker alternation constraint (Figure 2), whereas a total of 1,306 utterances of the lower half of the table were NOT evaluated with the speaker alternation constraint (Figure 3).]

when suffixes (for example, representing politeness) were added to those personal nouns, such as "会社員さん (kaishain-san)" (company employee + politeness) and "使用人くん (shiyonin-kun)" (servant + politeness), those personal nouns were also extracted because suffixes often help readers identify the speaker.

3.2 Dataset

We obtained 3,993 utterances as a consequence of the annotation, in which the speaker and personal nouns within the utterance were linked. The main characters were "Marie" (female), and "Kyuros" (male). Of the 3,993 utterances, 1,178 were by "Marie" (female) and 1,030 by "Kyuros" (male), and the remaining 1,785 utterances were by minor characters [5]. In this research, the performance of the speaker identification was exclusively evaluated on the 2,208 utterances by "Marie" (female) and "Kyuros" (male). Due to the gender classification model's constraints on input token size, only 32 tokens from the beginning of the utterance were used for classification.

Table 3 shows the dataset statistics, focusing on utterances by "Marie" (female) or "Kyuros" (male). The upper half of the table shows the number of dialogues and utterances consisting only of utterances by "Marie" (female) or "Kyuros" (male), whereas the lower half shows that of dialogues including characters other than "Marie" (female) or "Kyuros" (male) and the number of isolated single utterances immediately preceded/followed by narratives. The upper half of Table 3 shows that the speaker alternation constraint (described in section 5) was satisfied for almost all the dialogues consisting only of utterances by "Marie" (female) and "Kyuros" (male). Those consecutive utterances to which the speaker alternation constraint could be applied were those immediately preceding or immediately following another utterance, where isolated single utterances immediately preceded/followed by narratives were excluded.

Table 4 shows the statistics for the utterances subject to the constraint on personal nouns within quotes (described in section 6). Utterances that meet the constraint on personal nouns within quotes

[5] Of the 1,785 utterances, 1,277 of the dialogues were with "Marie" (female) or "Kyuros" (male) and the remaining 508 utterances were from other dialogues. (shown in the lower half of Table 3)

Character	Proportion
“Kyuros” (male)	49.2% (507/1,030)
“Marie” (female)	39.8% (469/1,178)
Total	44.2% (976/2,208)

Table 4: Applicability of the Constraint on Personal Nouns Within Quotes

were those that contained at least one personal noun within the quote.

4 Gender Classification Model by BERT

In this section, we describe how to construct the gender classification model, which is based on a fine-tuned BERT-based gender-specific language model (Devlin et al., 2019). Different linguistic styles in the Japanese language convey the speaker’s gender (Murai, 2018). Therefore, the linguistic style of the utterance can be used to determine the gender of the speaker. Our gender classification model takes an utterance as input and outputs the classification probability of the speaker’s gender. This model is similar to the persona speaker probability model of Miyazaki et al. (2021) used for filtering out inappropriate utterances with respect to the given persona. The persona speaker probability model was trained by collecting utterances from the set of personas. In contrast to their method, we used utterances from a large number of novels other than the target novel as the training data to train the classification model.

4.1 First-Person Pronouns in the Japanese Language

There are numerous types of personal pronouns used in the Japanese language. Personal pronouns in Japanese can approximately indicate various attributes, such as gender, age, temperament, and characters’ social status. Significantly, the first-person pronouns “俺 (ore)” and “私 (watashi)” are remarkably contrasting, with “俺 (ore)” being primarily used by males and “私 (watashi)” being primarily used by females. It is observed that words commonly used by males frequently appear in utterances containing “俺 (ore)”, and similarly, words commonly used by females appear in utterances containing “私 (watashi)”. Considering this, we constructed a dataset of utterances that included linguistic expressions specific to genders, containing the first-person pronouns “俺 (ore)” or “私 (watashi)” from novels.

4.2 Collecting Quotes for Training the Gender Classification Model

We used a novel posting site called “小説家になろう” (Aim to be a novelist)[6] to gather utterances containing the first-person pronouns “俺 (ore)” and “私 (watashi)”. We selected 2,000 novels and gathered their text. We next applied the morphological analysis to the text using Sudachi[7]. Quotes in Japanese text were represented “「” at the start and “」” at the end, as present below:

> 「おはよう、マリー。」
> (“Good morning, Marie.”)

Therefore, to extract quotes from the Japanese text, we extracted string sequences starting with “「” and ending with “」”. Furthermore, we selected quotes that included the first-person pronouns “俺 (ore)” or “私 (watashi)”. At this stage, we obtained 38,192 quotes that included “俺 (ore)” and 52,052 quotes that included “私 (watashi)”. Then, quotes composed of multiple sentences were decomposed according to the symbols “。”,“?”, and “!” and were regarded as multiple quotes[8]. By accumulating each of those multiple quotes, we obtained 83,571 utterances that had linguistic expressions closely related to “俺 (ore)” (typical for males) and 118,997 utterances that had linguistic expressions closely related to “私 (watashi)” (typical for females).

4.3 Training the Gender Classification Model

In this research, we used a pre-trained BERT (Devlin et al., 2019) model for gender classification. Specifically, we used Tohoku University’s Japanese version of BERT-base[9], which is trained on Japanese Wikipedia[10]. This model consists of 12 layers,

[6] `https://syosetu.com/`

[7] `https://github.com/WorksApplications/Sudachi`

[8] After decomposing the original single quote into multiple quotes, it can happen that those constituent quotes may not include the first-person pronouns “俺 (ore)” or “私 (watashi)”.

[9] `https://github.com/cl-tohoku/bert-japanese`

[10] We also evaluated a base-sized Japanese RoBERTa model (Liu et al., 2019) of `https://huggingface.`

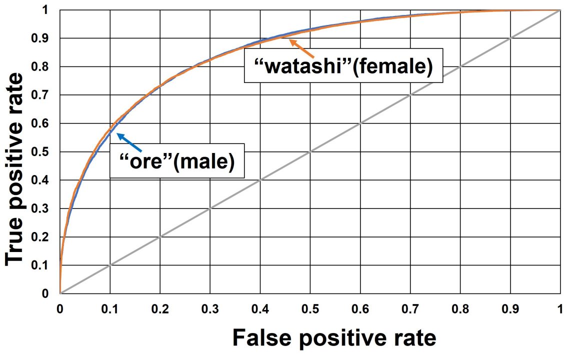


Figure 1: ROC Curves of the Gender Classification Model for the Test Data (%) [The set typical for males denoted as "ore" (male) and the set typical for females denoted as "watashi" (female). The overall test data accuracy is 77.1% and the AUC is 0.851 for both curves]

768 dimensions of hidden states, and 12 attention heads. The first-person pronouns "俺 (ore)" and "私 (watashi)" of the input sentence were replaced with the [MASK] token in both training and testing[11]. For a total of 83,571 utterances that had linguistic expressions closely related to "俺 (ore)" (typical for male) and 118,997 utterances that have linguistic expressions closely related to "私 (watashi)" (typical for female), the ratio of training, validation and test data was 6:2:2. The model with the minimum validation loss was evaluated against the test data.

4.4 Evaluating the Gender Classification Model

Figure 1 shows the ROC-curves of our gender classification model for male (denoted as "ore" (male), i.e., utterances that have linguistic expressions closely related to "俺 (ore)") and for female (denoted as "watashi" (female), i.e., utterances that have linguistic expressions closely related to "私 (watashi)") for the test data, where the accuracy of the overall test data is 77.1% and each of the two curves has AUC of 0.851. Table 5 shows examples of probabilities by the gender classification model. The utterances including "だぜ (daze)" are identified as related to the first-person pronoun "俺 (ore)" (male). In contrast, the utterances including "よ (yo)" are identified as related to the first-person pronoun "私 (watashi)" (female), which is consistent with the observation of Murai (2018). In addition, the utterance with "お前 (omae)" which is used primarily by males, and that with "あなた (anata)" which is used primarily by females, are also appropriately identified by our gender classification model.

5 Speaker Alternation Constraint

In consecutive utterances in a novel, two speakers usually take turns alternately. This observation is represented in this section by the "speaker alternation constraint", where we restricted the turns of the speakers to constantly alternating. We also made a minor modification to the speaker alternation constraint, where we defined a dialogue as a sequence of utterances with no intervening narratives. This modification stems from the observation that when narratives interrupt in the middle of utterances, the same speaker is often consecutive. He et al. (2013) proposed integrated classifier features for speaker identification with the speaker alternation constraint. In contrast to He et al. (2013), we introduce the speaker alternation constraint as the hard constraint in this research, so that the results of speaker identification strictly follow the constraint.

This constraint was satisfied for almost all the dialogues consisting only of utterances by "Marie" (female) and "Kyuros" (male), as shown in the upper half of Table 3. More specifically, as mentioned in section 3.2, as the target speakers in the novel for evaluation, we considered only the two major characters of the romance novel, "Marie" (female) and "Kyuros" (male). In terms of the speaker alternation constraint, we only considered those two major characters.

Let A and B be the two speakers for which we consider the speaker alternation constraint. Given an utterance sequence $(U_1, U_2, \ldots)$ and suppose that $(S_1, S_2 \ldots)$ is the speaker sequence of $(U_1, U_2, \ldots)$, then under the speaker alternation constraint, we considered only the following two types of speaker

co/rinna/japanese-roberta-base in the same task, where the performance was almost similar to that of BERT (Devlin et al., 2019).

[11] The performance with [MASK] token replacement was superior to than that without [MASK] token replacement for both training and testing.

Example Sentences	Probability of the Model	
	“俺 (ore)” (typically for male)	“私 (watashi)” (typically for female)
俺のものだぜ ore no mono **da** **ze** (It’s mine)	**0.965**	0.035
私のものよ watashi no mono **yo** (It’s mine)	0.005	**0.995**
お前の名前は? **omae** no namae wa ? (What is your name?)	**0.769**	0.231
あなたの名前は? **anata** no namae wa ? (What is your name?)	0.083	**0.917**

Table 5: Examples of the Probabilities by the Gender Classification Model

Personal Nouns (Japanese/pronunciation/English)	# of times the quotes including the personal noun occurrences judged by the gender classification model		Total
	Marie (female)	Kyuros (male)	
わたし/“watashi”/I (only used by Marie)	215 (100.0%)	0 (0.0%)	215
マリー/“mari”/Marie (mostly used by Kyuros)	19 (8.9%)	194 (91.1%)	213
俺/“ore”/I (only used by Kyuros)	3 (1.7%)	174 (98.3%)	177
君/“kimi”/you (only used by Kyuros)	40 (33.1%)	81 (66.9%)	121
ミオ/“mio”/Mio (used by both Marie and Kyuros)	41 (54.7%)	34 (45.3%)	75

Table 6: Examples of the Personal Nouns and # of Times their Occurrences Judged by the Gender Classification Model

sequences:

$$(S_1, S_2, S_3, \ldots) = (A, B, A, \ldots)$$
$$(S_1, S_2, S_3, \ldots) = (B, A, B, \ldots)$$

Here, the probability of the speaker sequence of the utterance sequence $(U_1, U_2, \ldots, U_n)$ being $(S_1, S_2, \ldots, S_n)$ by the gender classification model P_g is given as below:

$$P_g((U_1, U_2, \ldots), (S_1, S_2, \ldots)) = \prod_{i=1}^{i=n} P_g(U_i, S_i)$$

Therefore, we compared the two probabilities below:

$$P_g((U_1, U_2, \ldots), (A, B, A, \ldots))$$
$$P_g((U_1, U_2, \ldots), (B, A, B, \ldots))$$

As the result of the gender classification model and the speaker alternation constraint, the speaker sequence with the higher probability was chosen.

In the evaluation, the speaker alternation constraint was applied to the 902 utterances of dialogues only by “Marie” (female) and “Kyuros” (male) in the upper half of Table 3.

6 Constraint on Personal Nouns Within Quotes

Personal nouns, as well as Japanese suffixes representing politeness such as “君 (kun)” and “さん (san)” within utterances, often help readers in identifying the speaker. As shown in Table 6, the first-person pronoun “わたし (watashi)” was only used by “Marie” (female), whereas the first-person pronoun “俺 (ore)” was only used by “Kyuros” (male). Furthermore, the second-person pronoun “君 (kimi)” was only used by “Kyuros” (male). However, as in Table 6, according to the gender classification model, 33% of the utterances including “君 (kimi)” were incorrectly judged as uttered by “Marie” (female). To correct those 33% error cases, we introduced the score $Q_{G+PN}(U, S)$ of the speaker of the utterance U to be S by the gender classification model adjusted by the constraint of the personal nouns within quotes as shown below:

$$Q_{G+PN}(U, S) \;=\; P_g(U, S) + \alpha \sum_{k=1}^{m} r_g(n_k, S) \quad (1)$$

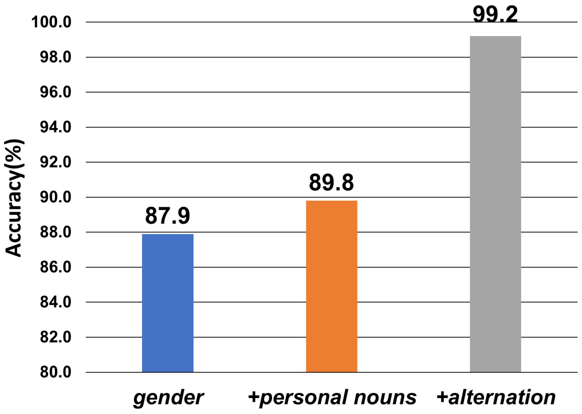


Figure 2: Evaluation Results for 902 Utterances in the Upper Half of Table 3 Consisting of Utterances Only by Marie (female) and Kyuros (male) (*gender*: gender classification model, +*personal nouns*: gender classification model + the constraint on personal nouns within quotes, +*alternation*: gender classification model + the speaker alternation constraint)

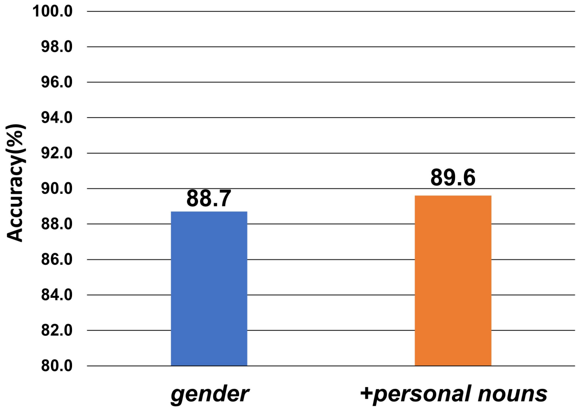


Figure 3: Evaluation Results for 1,306 Utterances [by Marie (female) and Kyuros (male)] in the Lower Half of Table 3 (*gender*: gender classification model, +*personal nouns*: gender classification model + the constraint on personal nouns within quotes)

where P_g is the probability of the gender classification model, $n_1, n_2, \dots n_m$ represent the m kinds of personal nouns within the utterance U, $r_g(n_k, S)$ represents the proportion of utterances classified by the gender classification model as speaker S out of the total utterances containing the personal noun n_k, and $\alpha = 0.895$ is a hyper-parameter to be optimized with accuracy through 10-fold cross-validation. The intuitive motivation of the formula (1) is to revise the probability $P_g(U, S)$ of the gender classification model by adding the proportion $r_g(n_k, S)$ of utterances classified by the gender classification model for all of the personal nouns $n_1, n_2, \dots n_m$ within the utterance U.

For example, suppose the utterance U includes the second-person pronoun “君 (kimi)”. Then, the score $Q_{G+PN}(U, S)$ for S =“Marie” (female) and S =“Kyuros” (male) can be represented as follows:

$$
\begin{aligned}
Q_{G+PN}(U, S = \text{“Marie” (female)}) &= P_g(U, S) \\
&\quad + \alpha r_g(\text{“君 (kimi)”}, S) \quad (r_g = 0.331) \\
Q_{G+PN}(U, S = \text{“Kyuros” (male)}) &= P_g(U, S) \\
&\quad + \alpha r_g(\text{“君 (kimi)”}, S) \quad (r_g = 0.669)
\end{aligned}
$$

Thus, even if in some cases, $P_g(U, S = \text{“Marie” (female)}) > P_g(U, S = \text{“Kyuros” (male)})$, by adding $\alpha r_g(\text{“君 (kimi)”}, S = \text{“Kyuros” (male)})$ $(= 0.895 \times 0.669)$ to $P_g(U, S = \text{“Kyuros” (male)})$, finally $Q_{G+PN}(U, S = \text{“Kyuros” (male)})$ becomes greater than $Q_{G+PN}(U, S\text{=“Marie” (female)})$, then the classification error by the gender classification model can be recovered.

7 Evaluation

We used the dataset of utterances from two major characters described in section 3 for the evaluation experiment. When the gender classification model categorized an utterance as male, it was labeled as “Kyuros” (male), and when it classified an utterance as female, it was labeled as “Marie” (female).

Figure 2 shows the accuracy of the gender classification model (denoted as *gender*), the gender classification model + the constraint on personal nouns within quotes (denoted as +*personal nouns*), and the gender classification model + the speaker alternation constraint (denoted as +*alternation*) for 902 utterances in the upper half of Table 3 consisting of utterances only by “Marie” (female) and “Kyuros” (male). The gender classification model that was trained using an automatically collected dataset achieved an accuracy of 87.9% for the classification of the two main characters of the target novel. Both the speaker alternation constraint and the constraint on personal nouns within quotes were found to improve the classification performance. The speaker alternation constraint was satisfied for

Additional Constraints	Applicable Utterances	Classification Error before Application	After Application	
			Improved	Damaged
Speaker Alternation Constraint	902	109 (12.1%)	106 (11.8%)	4 (0.4%)
Constraint on Personal Nouns within Quotes	976	82 (8.4%)	31 (3.2%)	5 (0.5%)

Table 7: Detailed Analysis of the Performance of Each Constraint

Example Sentences	Speaker	Probability of the Model		corrected by constraints
		Marie (female)	Kyuros (male)	
どういたしまして。これは余り do, itashi mashite. kore wa amari (You are welcome. This is a surplus.)	Kyuros	**0.812**	0.188	corrected by the speaker alternation constraint
さあ、マリー saa Marie (Here, Marie)	Kyuros	**0.512**	0.487	corrected by the constraint on personal nouns within quotes

Table 8: Examples of Classification Error by the Gender Classification Model and Corrected by the Speaker Alternation Constraint / the Constraint on Personal Nouns Within Quotes

almost all cases, as shown in the upper half of Table 3. Because of this high rate, the speaker alternation constraint achieved an almost perfect accuracy of 99.2%. However, the constraint on personal nouns within quotes slightly improved the accuracy of the gender classification model by just 1.9 points.

The evaluation results for 1,306 utterances by “Marie” (female) and “Kyuros” (male) in the lower half of Table 3 are also shown in Figure 3, where we did not apply the speaker alternation constraint because those dialogues included characters other than “Marie” (female) and “Kyuros” (male). Their evaluation results are mostly similar to those in Figure 2, where the constraint on personal nouns within quotes slightly improved the accuracy of the gender classification model.

Table 7 shows a detailed analysis of each constraint’s performance. As shown in Tables 3 and 4, the proportion of utterances to which each constraint is applicable was not insignificant. However, the performance of the gender classification model was relatively high. Table 7 shows that the proportion of classification error before the application of each constraint was not particularly high. Therefore, the improvement that can be achieved by each constraint was also relatively insignificant.

Table 8 shows examples of classification errors by the gender classification model, where those errors were corrected by the speaker alternation constraint or the constraint on personal nouns within quotes. As in the first example, utterances containing honorifics that are used regardless of gender, such as sentence-final auxiliary verbs “です (desu)” and “ます (masu)”, have a higher probability of being classified as utterances by a female. The first example contains a conjugation form “まして (mashite)” of “ます (masu)” and causes a classification error. This example is corrected by the speaker alternation constraint. Next, the second example consists of only a few words or lacks linguistic features that indicate gender and is incorrectly classified. This example, however, contains the personal noun “Marie” within the quote, which enables it to be correctly classified as an utterance by “Kyuros” (male).

8 Concluding Remarks

This research described a method for constructing a novel dataset by automatically collecting sentences having gender characteristics using first-person pronouns as a cue and a gender classification model trained on the dataset. This gender classification model shows high classification performance in the male and female speaker classification, indicating that gender-specific linguistic features contribute to the speaker identification task in Japanese novels. Additionally, we found incorporating personal nouns within the utterance and the preceding and following utterances increased the classification performance in Japanese novels, as in the case of English novels reported in previous studies (O’Keefe et al., 2012; Muzny et al., 2017). Although the performance of the gender classification model was relatively high, our task setting of speaker identification was limited to the two major characters of the romance novel. Our future work includes developing a speaker identification model that links all utterances to appropriate speakers using the classification probability by the gender classification model as one feature. It is also necessary to broaden the dataset to a more diverse set of contemporary novels.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21H00901.

References

Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. 2018. Unsupervised learning of style-sensitive word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–578, Melbourne, Australia, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 1013–1019.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ryo Ishii, Ryuichiro Higashinaka, Koh Mitsuda, Taichi Katayama, Masahiro Mizukami, Junji Tomita, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Yushi Aono. 2021. Methods for efficiently constructing text-dialogue-agent system using existing anime characters. *Journal of Information Processing*, 29:30–44.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository, arXiv:1907.11692*.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium, October-November. Association for Computational Linguistics.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 307–314, Shanghai, China, October.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2016. Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 319–328, Los Angeles, September. Association for Computational Linguistics.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 178–189, Singapore and Online, July. Association for Computational Linguistics.

Hajime Murai. 2018. Factor analysis of utterances in Japanese fiction-writing based on BCCWJ speaker information corpus. *Advances in Human-Computer Interaction*, 2018:1–9, 11.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain, April. Association for Computational Linguistics.

Timothy O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, Jeju Island, Korea, July. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *Computing Research Repository, arXiv:1512.05742*.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. *Computing Research Repository, arXiv:2109.05217*.

Makoto Yamazaki, Yumi Miyazaki, and Wakako Kashino. 2018. Annotation and quantitative analysis of speaker information in novel conversation sentences in Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chak Yan Yeung and John Lee. 2017. Identifying speakers and listeners of quoted speech in literary works. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 325–329, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July. Association for Computational Linguistics.

Developing and Evaluating a Dataset for How-to Tip Machine Reading at Scale

Fuzhu Zhu, Shuting Bai, Tingxuan Li, Takehito Utsuro
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba,
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
{s2220804,s2020817,s2120816}_@_s.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

In this paper, we focus on the task of machine reading at scale within how-to tip machine reading comprehension (MRC). We propose a method for developing a context dataset using how-to tip websites on the Internet as information sources. This shows that the proposed method can easily create a context dataset containing thousands of context sets. Furthermore, this paper uses a method for retrieving the context from the developed context dataset, which contains the answer of the question. It applies to the MRC model. Specifically, we use three models based on TF-IDF and BERT (TF-IDF, BERT, and TF-IDF+BERT) as our retrieval models. Meanwhile, the BERT model served as our MRC model. We apply the retrieval model and the MRC model to the context dataset after combining them. Evaluation results show that the TF-IDF+BERT model outperforms the other two models when tested against the context dataset.

1 Introduction

In natural language processing, machine reading comprehension (MRC) tasks are formulated to extract the answer to a question from a context within a few question sentences and contexts expressed in natural language. MRC tasks can be divided into two categories based on the two types of answers. Factoid MRC tasks aim at having the answer to factoids such as proper nouns and numbers, where the answer is usually unique, short and simple. Conversely, nonfactoid MRC tasks aim to obtain an answer about nonfactoid such as explanation, reason and how-to tip, where there are usually multiple options and the answer is frequently a full sentence, rahter than a word or phrase. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016)is one of the most well-known QA datasets and benchmark tests among factoid MRC related to Wikipedia articles and news articles. Additionally, it is acknowledged that recent deep learning models (for example, BERT (Devlin et al., 2019)) trained with SQuAD achieved fairly high performance[1]. However, some research cases are known for nonfactoid MRC. They include MS MARCO (Nguyen et al., 2016), which has been developed using Bing's search logs and passages of retrieved web pages; DuReader (He et al., 2018), which has been developed using Baidu Search; Baidu Zhidao, a Chinese community-based QA site; and the NarrativeQA (Kočiský et al., 2018) dataset (in English), which contains questions created by editors based on summaries of movie scripts and books. They also include Soleimani et al. (2021), Dulceanu et al. (2018), and Cohen et al. (2018). Among those working on nonfactoid MRC, the case of MRC of Japanese how-to tip QAs (Chen et al., 2020) selected the how-to tip websites that are posted on the Internet and chose the column pages on how-to tip websites as information sources to collect how-to tip QA examples for training and testing. It has also been shown that the how-to tip MRC model with specific performance can be developed.

Figure 1 shows the how-to tip MRC model (Chen et al., 2020). The how-to tip MRC model and the

[1] https://rajpurkar.github.io/SQuAD-explorer/

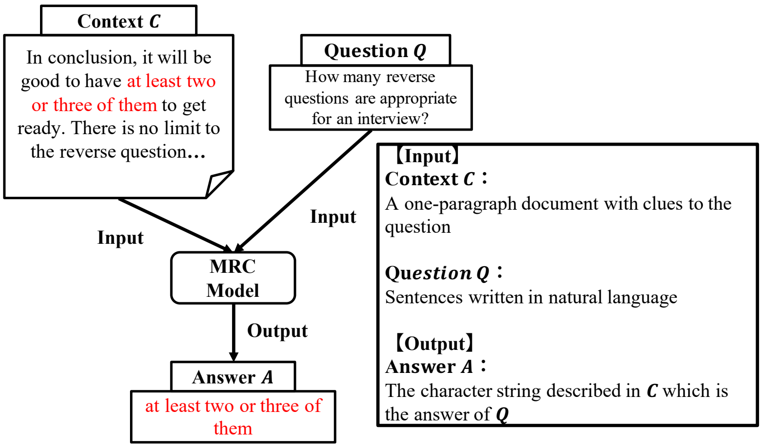

Context *C*
In conclusion, it will be good to have at least two or three of them to get ready. There is no limit to the reverse question…
Question *Q*
How many reverse questions are appropriate for an interview?
Input
Input
MRC Model
Output
Answer *A*
at least two or three of them
【Input】
Context *C*：
A one-paragraph document with clues to the question
Question *Q*：
Sentences written in natural language
【Output】
Answer *A*：
The character string described in *C* which is the answer of *Q*

Figure 1: The framework of how-to tip MRC model

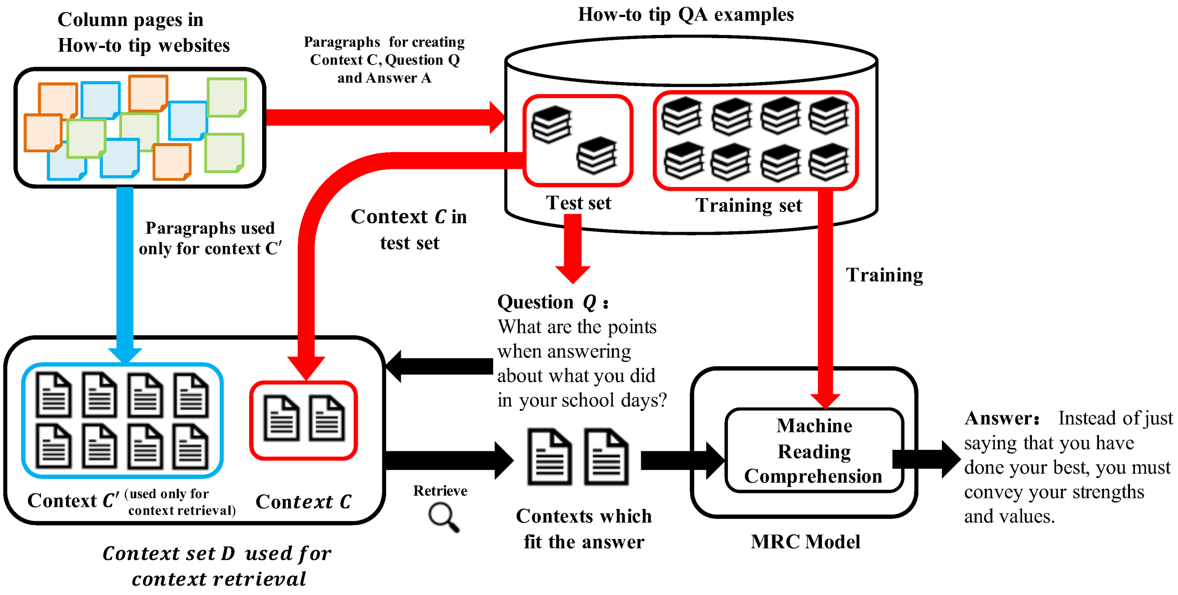

Column pages in How-to tip websites
Paragraphs for creating Context C, Question Q and Answer A
How-to tip QA examples
Test set
Training set
Paragraphs used only for context C′
Context *C* in test set
Training
Question *Q*：
What are the points when answering about what you did in your school days?
Context *C*′ (used only for context retrieval)
Context *C*
Retrieve
Contexts which fit the answer
Machine Reading Comprehension
MRC Model
Answer： Instead of just saying that you have done your best, you must convey your strengths and values.
Context set D used for context retrieval

Figure 2: Developing a context dataset for how-to tip MRC by using column pages on how-to tip websites

framework of the typical MRC model, which contains a tuple of a context, a tip question, and an answer, can be represented as in this figure. Note that the answer is extracted only from the context. Therefore, in the situation where it is not given which context to be used, another framework called "machine reading at scale" (Chen et al., 2017) should be invented. In the framework of "machine reading at scale," it handles both information retrieval and MRC tasks. In its framework, the MRC model is applied to the set of candidate contexts retrieved by the information retrieval module. For example, in the information retrieval module, Chen et al. (2017) used the method of TF-IDF to collect the candidate contexts. As another example of "machine reading at scale," using the BERT (Devlin et al., 2019) model as part of the information retrieval model for machine reading at scale tasks has also been studied by Karpukhin et al. (2020). It is shown that using the BERT model as part of the information retrieval model, higher retrieval accuracy than the BM25 method can be achieved in several factoid MRC datasets. Moreover, it shows that the retrieval accuracy was further improved using the proposed retrieval model and the BM25 score together.

Based on that background, this paper applies the framework of "machine reading at scale" to how-to tip MRC. In this paper, we use three different types of retrieval models (context retrieval by TF-IDF (Chen et al., 2017), BERT model, and combining TF-IDF with the BERT model) and how-to tip MRC model (Chen et al., 2020) to how-to tip MRC tasks. Chen et al. (2020) chose the column pages in how-to tip websites as information sources to collect how-to tip QA examples as the training and test sets for the how-to tip MRC model. In this paper, we collect the contexts from the column pages that were not used to form the training and test sets of the how-to tip MRC model in Chen et al. (2020) as shown in the framework in Figure 2 and the example in Figure 3. Then, we use those collected contexts as the contexts C' (used only for context retrieval but not for the MRC model training) for context retrieval and how-to tip MRC task. In this paper, according to the procedures above, we finally develop a dataset for how-to tip machine reading at scale. As for the contexts C', thousands of them are collected.

2 A Dataset for How-to Tip Machine Reading at Scale

In this section, we will introduce how to collect the context C' used only for context retrieval in Figure 2 and how to develop a dataset for how-to tip machine reading at scale.

Japanese how-to tip websites were selected from six types of topics[2] (which are "job hunting," "marriage," "apartment," "hay fever," "dentist," and "food poisoning") by Chen et al. (2020). After that, they collected column pages from the how-to tip websites[3]. Finally, a maximum of five paragraphs were selected from each column page, and they used them as contexts for constructing answerable/unanswerable how-to tip QA examples. An answerable how-to tip QA example contains Context C, Question Q, and Answer A, whereas an unanswerable how-to tip QA example contains Context C, Question Q, and Answer $A' = \langle \text{null} \rangle$[4].

Considering the above procedure of Chen et al. (2020), this section shows how we collect the context C' used only for context retrieval in Figure 2. More specifically, as shown in the example of Figure 3, within the column page used by Chen et al. (2020), we do not use the maximum five paragraphs selected by Chen et al. (2020) (as shown in the red boxes). Still, we use those other than the maximum five paragraphs (as shown in the blue boxes). We also carefully examine the context dataset of Chen et al. (2020), which was developed manually by selecting the paragraph used, and we follow the standards below to select the candidate paragraphs efficiently:[5][6]

(i) Based on the restriction when applying the MRC models by BERT (Devlin et al., 2019), the upper bound of the number of morphemes within a paragraph is set to 290[7].

(ii) The lower bound of the number of characters in a paragraph is 30.

(iii) Any URL is excluded from the paragraph.

(iv) Any email addresses were excluded from the paragraph.

Table 1 shows the number of web pages used for each topic ("job hunting," "marriage," "apartment," "hay fever," "dentist," and "food poisoning"). It also shows the number of contexts used for constructing how-to tip QA examples and the number of contexts used only for context retrieval[8]. Figure 3 shows how

[2] The specific term used in Chen et al. (2020) is "query focus," rather than "topic." The notion of *query focus* is a keyword used for every search request related to a specific subject. In this paper, however, for simplicity, we use the term "topic" in stead of "query focus."

[3] The detailed procedure of collecting pages from the how-to tip websites is stated by Chen et al. (2020).

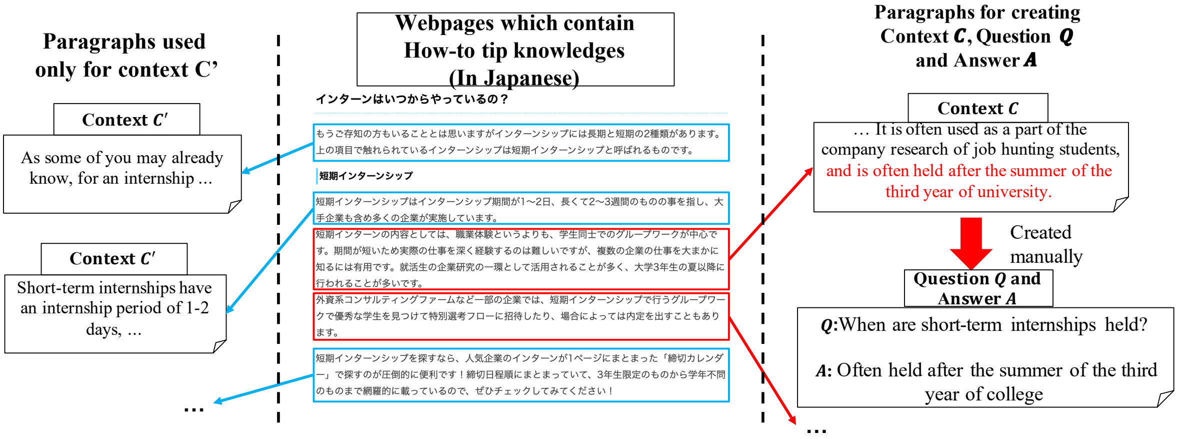
[4] Both SQuAD1.1-type answerable and SQuAD2.0-type unanswerable QA examples were created from the same column page (Chen et al., 2020).

[5] The standard (i) is simply for satisfying the requirement when applying the MRC models by BERT. Conversely, the standards (ii), (iii), and (iv) are for avoiding paragraphs that do not have sufficient how-to tip knowledge. These standards are also for avoiding the task of filtering out the context C' used only for the context retrieval to be too easy.

[6] One of the authors of the paper performed the procedure of manual selection.

[7] In the experiments and evaluation of the retrieval module throughout this paper, MeCab (`https://taku910.github.io/mecab/`) and mecab-ipadic-NEologd (`https://github.com/ neologd/mecab-ipadic-neologd`) are used in Japanese morphological analysis.

[8] Those data shown in Table 1 (except for "The number of contexts only for retrieval"), Table 2, and Table 3 are essentially the same as those reported in Chen et al. (2020), but we


Paragraphs used
only for context C'
Context C'
As some of you may already know, for an internship ...
Context C'
Short-term internships have an internship period of 1-2 days, ...
...
Webpages which contain
How-to tip knowledges
(In Japanese)
Paragraphs for creating
Context C, Question Q
and Answer A
Context C
... It is often used as a part of the company research of job hunting students, and is often held after the summer of the third year of university.
Created manually
Question Q and Answer A
Q:When are short-term internships held?
A: Often held after the summer of the third year of college
...

インターンはいつからやっているの？

もうご存知の方もいることとは思いますがインターンシップには長期と短期の2種類があります。上の項目で触れられているインターンシップは短期インターンシップと呼ばれるものです。

短期インターンシップ

短期インターンシップはインターンシップ期間が1〜2日、長くて2〜3週間のものの事を指し、大手企業も含め多くの企業が実施しています。

短期インターンの内容としては、職業体験というよりも、学生同士でのグループワークが中心です。期間が短いため実際の仕事を深く経験するのは難しいですが、複数の企業の仕事を大まかに知るには有用です。就活生の企業研究の一環として活用されることが多く、大学3年生の夏以降に行われることが多いです。

外資系コンサルティングファームなど一部の企業では、短期インターンシップで行うグループワークで優秀な学生を見つけて特別選考フローに招待したり、場合によっては内定を出すこともあります。

短期インターンシップを探すなら、人気企業のインターンが1ページにまとまった「締切カレンダー」で探すのが圧倒的に便利です！締切日程順にまとまっていて、3年生限定のものから学年不問のものまで網羅的に載っているので、ぜひチェックしてみてください！

Figure 3: Using a column page to collect contexts for creating how-to tip QA examples (from: "When does job hunting begin?" (in Japanese) (`https://internshipguide.jp/columns/view/shukatsu_sched_1`))

Table 1: The number of used web pages and the number of collected contexts

Topic	Number of used web pages	The number of contexts for QA examples		The number of contexts only for retrieval
		Training examples	Test examples	
job hunting	293	1,478	98	4,675
marriage	182	1,386	98	2,868
apartment	50	—	100	491
hay fever dentist food poisoning	51	—	100	962

to collect contexts from a column page. Based on the procedures above, as shown in Figure 2, the context set D used for context retrieval consists of the context set C' used only for context retrieval and the context set C of the test examples of how-to tip QA examples.

3 BERT Retrieval Model

This section describes the structure of the BERT retrieval model devdeloped based on Karpukhin et al. (2020), the training method, and the retrieval procedure.

This BERT retrieval model uses two independent BERT models (Devlin et al., 2019)[9] as a question encoder E_q and a context encoder (in Karpukhin et al. (2020), passage encoder) E_c. The BERT model is applied to each input question Q and context C and the representations of the output CLS tokens are used as the representations $E_q(Q)$ and $E_c(C)$ of question Q and context C. The cosine similarity of the following equation is used as the similarity between the encoded Q and C.

$$sim(Q, C) = \frac{E_q(Q) \cdot E_c(C)}{\parallel E_q(Q) \parallel \ \parallel E_c(C) \parallel} \qquad (1)$$

In the process of training the model, for $i = 1, \ldots, m$, a set of the question Q_i, one relevant (positive) context C_i^+ that contains the reference answer and n irrelevant (negative) contexts C_i^- that do not contain the reference answer, is used as a training instance.

$$(Q_i, C_i^+, C_{i,1}^-, \ldots, C_{i,n}^-) \qquad (2)$$

and m sets of such a tuple are collected as a set T of training data.

$$T = \left\{ (Q_i, C_i^+, C_{i,1}^-, \ldots, C_{i,n}^-) \middle| i = 1, \ldots, m \right\}$$

We optimize the loss function below that is the neg-

increase their numbers through annotation to additional data. The reason why the number of examples for "apartment," "hay fever," "dentist," and "food poisoning" is less than those of "job hunting" and "marriage" is simply that the annotation procedure had started from "job hunting" and "marriage." It is quite possible to collect the same number of examples for each topic "apartment," "hay fever," "dentist," and "food poisoning."

[9] The BERT retrieval model was implemented using the HuggingFace version (`https://github.com/huggingface/ transformers`). A multilingual cased model was adopted as the pre-training model.

ative log likelihood of the positive context:

$$L(Q_i, C_i^+, C_{i,1}^-, \ldots, C_{i,n}^-) = \\ - log \frac{e^{sim(Q_i, C_i^+)}}{e^{sim(Q_i, C_i^+)} + \sum_{j=1}^{n} (e^{sim(Q_i, C_{i,j}^-)})} \quad (3)$$

Furthermore, to create the training dataset of a question Q and a context C that contains the reference answer above, we follow the strategy of “in-batch negatives” of Karpukhin et al. (2020). In this strategy, assume that we have B questions in a mini-batch and each one is associated with a relevant context. Roughly speaking, for each question Q_i in a mini-batch, there exist $B-1$ contexts, each of which is the relevant context of one of other $B-1$ questions in the same mini-batch. However, for the question Q_i, each of those $B-1$ contexts can be regarded as an irrelevant context. With this strategy, it enables us to create B training instances in each batch, where there are $B-1$ negative contexts for each question. This strategy is known as effective for boosting the number of training examples.

When we retrieve the contexts, the fine-tuned BERT model is used to pre-encode the contexts used for context retrieval, where the contexts are indexed using FAISS (Johnson et al., 2021) offline. For each question, the Top n contexts are output as retrieval results under the similarity scale of the formula (1).

4 Evaluation

4.1 The Dataset

Table 1 shows the number of web pages and the number of contexts used for creating how-to tip QA examples, as well as the number of contexts used only for context retrieval in the evaluation. Table 2 also shows the number of questions in how-to tip QA examples and Table 3 shows the number of how-to tip QA examples and factoid QA examples, respectively.

4.2 Evaluation Procedure

We use the following three types of context retrieval models to evaluate our dataset.

(i) TF-IDF model.

(ii) BERT retrieval model.

(iii) “TF-IDF+BERT” model, which takes the sum of (i) and (ii) scores.

Table 2: Number of questions related to how-to tip

topic	For creating Training set	For creating Test set
job hunting	795	50
marriage	799	49
apartment	—	50
others	—	49

Table 3: The number of QA examples

(a) factoid QA examples

training/test	The number of sets of context，question and answer (answerable/unanswerable)
training	$27,427/28,742$
test	$50/50$

(b) how-to tip QA examples

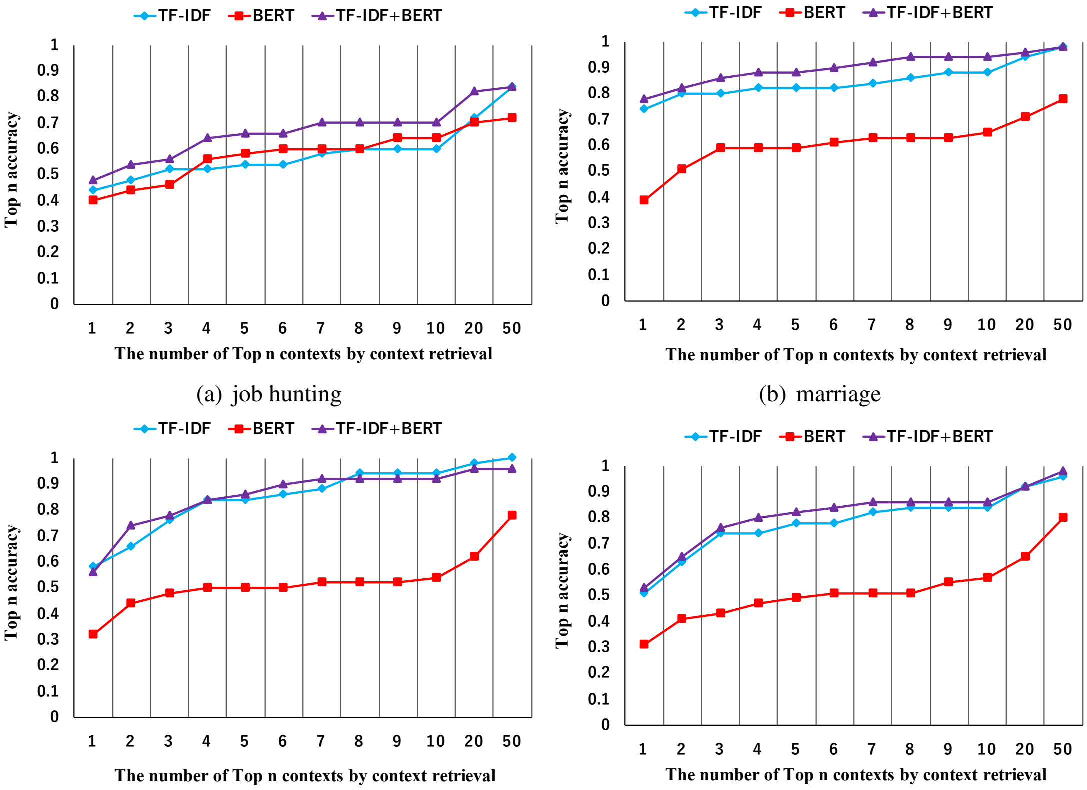
topic	The number of sets of context, question and answer (answerable/unanswerable)	
	Training set	Test set
job hunting	$807/807$	$50/50$
marriage	$807/807$	$50/50$
apartment	—	$50/50$
others	—	$50/50$

For (i), to build the TF-IDF (Chen et al., 2017) model[10], we add a stop word list in Japanese-SlothLib[11]. For each context set of the topics of “job hunting,” “marriage,” “apartment,” and the mixture of “hay fever,” “dentist,” and “food poisoning,” one TF-IDF model is built.

As described in Section 3, for (ii), we use the set of the pairs of question Q and the context C that contains the reference answer as the training data. The numbers of the set of the question Q, the context C, and the answer A are as shown in Table 3(b), where we use only the answerable training data for the topics “job hunting” and “marriage” and fine-tune the BERT retrieval model.

[10] https://github.com/facebookresearch/DrQA

[11] http://svn.sourceforge.jp/svnroot/slothlib/



(a) job hunting

(b) marriage

(c) apartment

(d) hay fever / dentist / food poisoning

Figure 4: Evaluation results of the three types of context retrieval models (with top n accuracy of the retrieved contexts)

For (iii), we use the inner product of the TF-IDF models feature vector of the question Q and the context C as the score $S_T(Q, C)$ of the TF-IDF model and use the cosine similarity between Q and C that are encoded by the BERT retrieval model as the score $S_B(Q, C)$. For one question Q_i, suppose that $S_T(Q_i, C_j)(j = 1, \dots, n)$ are the scores for the n candidate contexts[12]; the following equation gives the score $S_{T+B}(Q_i, C_j)$ of the "TF-IDF+BERT" model, which is the sum of the scores of (i) and (ii):

$$S_{T+B}(Q_i, C_j) = S_T(Q_i, C_j) + S_B(Q_i, C_j) \quad (4)$$

Based on S_{T+B}, we rank the candidate contexts, and the top k ($k = 1, \dots, n$) contexts are output as the results.

Meanwhile, the following three types of QA examples are used to fine-tune the BERT (Devlin et al., 2019) MRC model.

(i) Factoid QA examples (the training examples are shown in Table 3(a)).

(ii) How-to tip QA examples of "job hunting" and "marriage" (the training examples are shown in Table 3(b)).

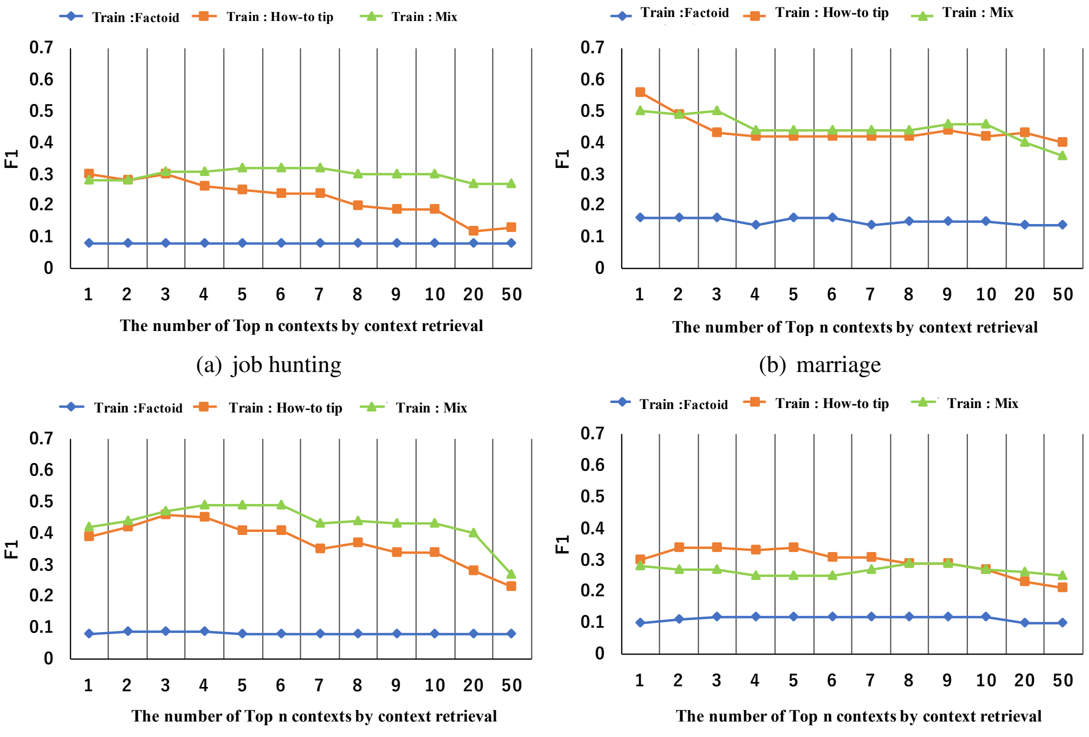
(iii) A mixture of both (i) and (ii).

As the version of the BERT implementation, which can handle a text in Japanese, the TensorFlow version[13] and the Multilingual Cased model[14] were used as the pre-trained model. Before applying BERT modules, MeCab was applied with IPAdic dictionary, and the Japanese text was segmented into a morpheme sequence. Then, within the BERT fine-tuning module, the WordPiece module with 110k shared WordPiece vocabulary was applied, and the Japanese text was further segmented into a subword

[12] The score $S_T(Q_i, C_j)(j = 1, \dots, n)$ is supposed to be normalized by the Min-Max method (minimum value is 0, whereas the maximum value is 1).

[13] `https://github.com/google-research/bert`

[14] Trained in 104 languages, available from `https://github.com/google-research/bert/blob/master/multilingual.md`.



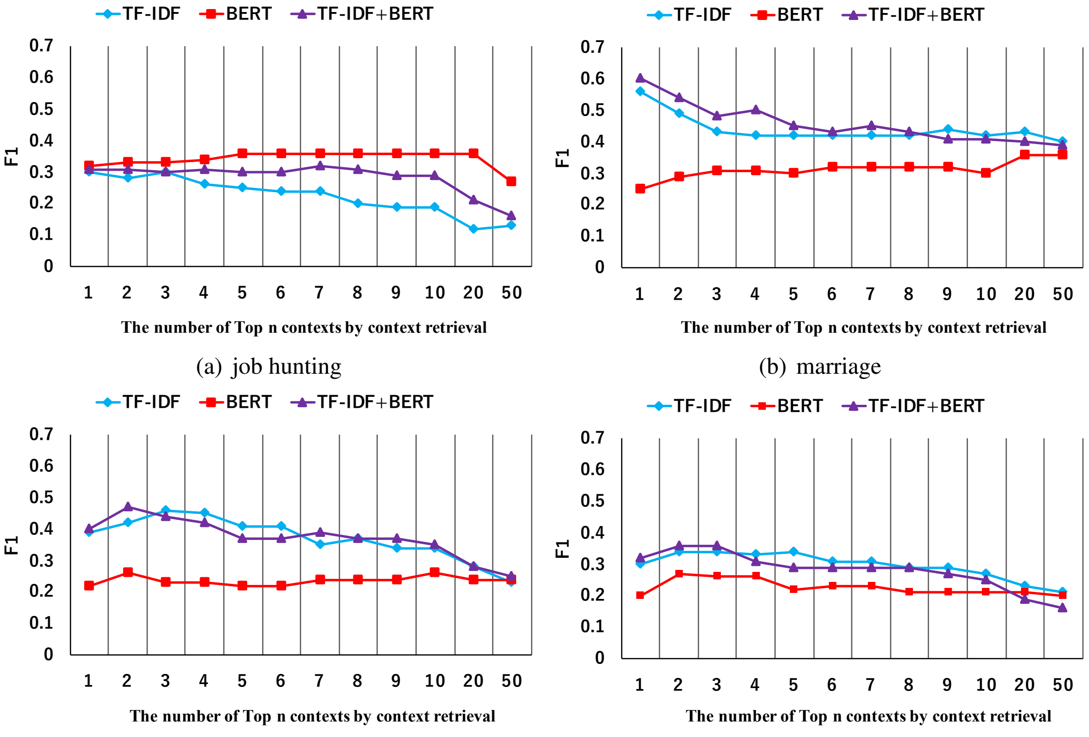
(a) job hunting

(b) marriage

(c) apartment

(d) hay fever / dentist / food poisoning

Figure 5: Evaluation results of machine reading at scale for three types of datasets used to fine-tune the BERT MRC model (with the TF-IDF model for context retrieval)



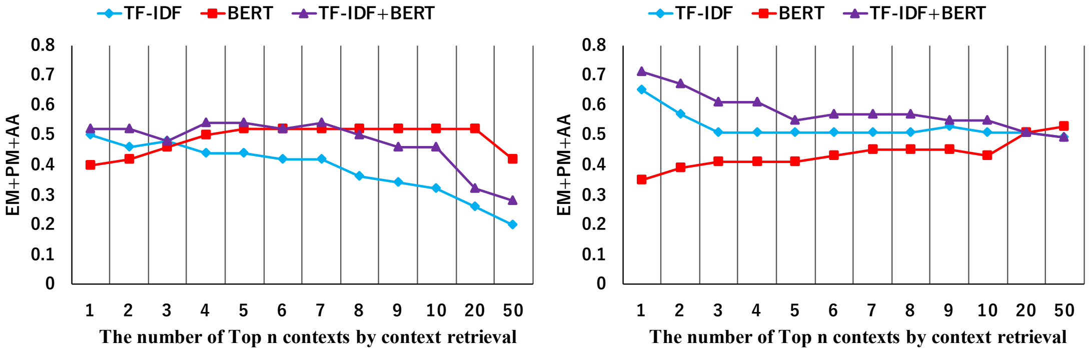
(a) job hunting

(b) marriage

(c) apartment

(d) hay fever / dentist / food poisoning

Figure 6: Evaluation results of the three types of context retrieval models (with the MRC model trained with how-to tip QA examples)



(a) job hunting

(b) marriage

Figure 7: Manual evaluation results of the three types of context retrieval models (with the MRC model trained with how-to tip QA examples)

unit sequence. Finally, the BERT fine-tuning module for MRC model[15] was applied.

The how-to tip MRC model is applied to top n ($n = 1, \ldots, 50$) retrieved contexts. Then, the answer with the highest score of the MRC model is chosen as the MRC model's output. Finally, we measure the F1 score which is calculated against the morpheme sequence of the reference answer.

In the manual evaluation[16], comparing the model's output and the reference answer, we evaluate the result manually according to the three evaluation criteria of "Exact Match" (EM), "Partial Match" (PM) and "Another Answer" (AA). We consider it the criterion for "Partial Match," when sufficient but partial information overlaps between the model's output and the reference answer. The criterion on "Another Answer," we consider it an answer when the condition "It is different from the reference answer, but contains enough information to answer the question" is satisfied. Then, we can calculate the ratio of the numbers of "Exact Match", "Partial Match" and "Another Answer" (EM+PM+AA).

4.3 Evaluation Result

Figure 4 shows the results of evaluating three types of context retrieval models in terms of top n retrieval accuracy, measured as the rate of queries for which the top n contexts include those with the reference answers. Figure 4(a) shows that the BERT retrieval model performs worse for cases other than "job hunting." This is mainly because, for the topics other than "job hunting," the queries for evaluation tend to include morphemes that appear in the contexts with the reference answers, which makes the TF-IDF model perform much better than the BERT retrieval model. For topic "job hunting," however, the queries for evaluation tend to include a relatively small number of morphemes that appear in the contexts with the reference answers, which happens to benefit the BERT retrieval model and makes it perform comparatively well with the TF-IDF model. By simply adding the scores of the two models, the "TF-IDF+BERT" model performs the best.

Figure 5 compares the three types of datasets used to fine-tune the BERT MRC model where the TF-IDF model is used for context retrieval. Similar to the evaluation results in Chen et al. (2020), also in the case of how-to tip MRC at scale in this paper, the performance of the model trained only by the factoid QA examples was significantly worse, whereas the one trained with the mixture of factoid + how-to tip QA examples performed the best. Overall, as the number of top n contexts increases, the model's performance tends to decrease on the contrary. This is simply because, as the number of top n contexts increases, not only those contexts with the reference answer, but also other contexts are included in the top n contexts, which damages the final MRC model results.

Figure 6 also compares the three types of context retrieval models, where the MRC model is trained with how-to tip QA examples[17]. Similarly, in Figure 4, the TF-IDF model performs well. Also, from

[15] `run_squad.py`, with the number of epochs of 2, batch size of 8, and learning rate of 0.00003.

[16] One of the author of the paper conducted a manual evaluation.

[17] The MRC model trained with the mixture of factoid + how-to tip QA examples shows almost a similar performance.

both Figure 5 and Figure 6, it can be seen that the MRC model trained with the topics of "job hunting" and "marriage" performs fairly well in how-to tip MRC on other topics such as "apartment," "hay fever," "dentist," and "food poisoning." From this result, it is sufficient to collect how-to tip QA examples only for one or two topics such as 'job hunting" and "marriage," and then fine-tune the MRC model, which applies to how-to tip MRC of any topic.

Finally, Figure 7 shows the manual evaluation result of the MRC model trained with how-to tip QA examples. Overall, the "TF-IDF+BERT" model performs the best in the evaluation of the performance of the MRC model for the topics of "job hunting" and "marriage." Compared with the automatic F1 results in Figure 6, it seems that the relative performance of the "TF-IDF+BERT" model improves simply because, by manual evaluation, certain nonliteral expressions within the "Another Answer" contribute to improving the performance of the "TF-IDF+BERT" model.

5 Related Work

Related studies of machine reading at scale, i.e., Chen et al. (2017), Karpukhin et al. (2020), Nishida et al. (2018), and Lee et al. (2019) investigated machine reading at scale in the context of factoid MRC. In Chen et al. (2017), machine reading at scale is realized by combining TF-IDF, which is used to realize context retrieval, and a neural MRC model using RNN. Karpukhin et al. (2020) used BERT (Devlin et al., 2019) for retrieval and then applied it to build a system for machine reading at scale. Moreover, in Nishida et al. (2018), machine reading at scale is realized via multi-task learning of information retrieval and MRC. Meanwhile, Lee et al. (2019) proposed an end-to-end framework for machine reading at scale that trains the retrieval and reading comprehension models.

In this paper, similar to Chen et al. (2017), TF-IDF model is used for the context retrieval part compared with those previous works, whereas another retrieval model (Karpukhin et al., 2020) by BERT is also investigated in this paper. For the part of reading comprehension, we use the BERT model instead. Combining these two parts, machine reading at scale is realized. Compared with that of Karpukhin et al. (2020), it is also important to note that we evaluate the performance change of the MRC model when the number of top n contexts increases, where it is observed that, in the case of our how-to tip QA examples, the optimal performance is around $n = 1$.

6 Conclusion

In this paper, we proposed a method to collect the contexts from the column pages that are not used to train the MRC model in Chen et al. (2020) and then use them to evaluate how-to tip machine reading at scale. Then, consequently, we developed a dataset that contains thousands of contexts for how-to tip machine reading at scale. Furthermore, we evaluated the three types of context retrieval models and showed that the "TF-IDF+BERT" model is the most effective. Future works include expanding the dataset as well as designing the evaluation procedure to be more reliable by introducing the notion of repeated trials and considering statistical measures such as variance (Dodge et al., 2020).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21H00901.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July. Association for Computational Linguistics.

Tengyang Chen, Hongyu Li, Miho Kasamatsu, Takehito Utsuro, and Yasuhide Kawada. 2020. Developing a how-to tip machine comprehension dataset and its evaluation in machine comprehension by BERT. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 26–35, Online, July. Association for Computational Linguistics.

Daniel Cohen, Liu Yang, and W. Bruce Croft. 2018. WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. In *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, pages 1165–1168, Ann Arbor, Michigan, U.S.A.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Finetuning pretrained language models: Weight initializations, data orders, and early stopping. *Computing Research Repository, arXiv:2002.06305*.

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy, July. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *Computing Research Repository*, arXiv:1611.09268.

Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 963–966.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255, Online, April. Association for Computational Linguistics.

Incorporating Linguistic Knowledge for Abstractive Multi-document Summarization

Congbo Ma, Wei Emma Zhang, Hu Wang, Shubham Gupta and Mingyu Guo
The University of Adelaide, Adelaide, Australia
{congbo.ma,wei.e.zhang,hu.wang,a1787223,mingyu.guo}@adelaide.edu.au

Abstract

Linguistic knowledge plays an important role in assisting models to learn informative representations that could help guide better natural language generation. In this paper, we develop a Transformer-based abstractive multi-document summarization method with linguistic-guided attention (LGA) mechanism for better representation learning. The proposed linguistic-guided attention mechanism can be seamlessly incorporated into multiple mainstream Transformer-based summarization models to improve the quality of the generated summaries. We develop the proposed method based on Flat Transformer (FT) and Hierarchical Transformer (HT), named ParsingSum-FT and ParsingSum-HT respectively. Empirical studies on both models demonstrate this simple but effective mechanism can help the models outperform existing Transformer-based methods on the benchmark datasets by a large margin. Extensive analyses examine different settings and configurations of the proposed model, providing a good reference to the text summarization community.

1 Introduction

Multi-document summarization (MDS) is a critical task in natural language processing aiming at generating a informative summary from a set of content-related documents. There are two types of summary generation: extractive summarization by selecting salient sentences from original texts directly and abstractive summarization to generate summaries by models from the understanding of the input contents (Ma et al., 2022). Under comparison, abstractive summarization is more challenging because it requires models to truly understand the input documents and generate corresponding summaries. With the development of deep learning techniques, neural network-based models that can help to capture high-quality latent features are widely applied in MDS (Dhakras et al., 2018; Alexander et al., 2019; Liu et al., 2019; Li et al., 2020; Han et al., 2020; Wen et al., 2021; Beltrachini et al., 2021).

Recently, Transformer (Ashish et al., 2017) shows outstanding performances in various natural language processing tasks, and it is also introduced into MDS (Alexander et al., 2019). Transformer has natural advantages for parallelization and could retain long-range relations between pairs of tokens among documents. Liu et al. (Liu et al., 2018) adopted a Transformer model to generate Wikipedia articles. The model selects top-K tokens and feeds them into the decoder-only sequence transduction. Built upon this work, Liu et al. (Liu et al., 2019) proposed a Hierarchical Transformer (HT) containing token-level and paragraph-level Transformer layers for cross-document relations capturing. Wen et al. (Wen et al., 2021) proposed a pre-train language model PRIMERA, using encoder-decoder transformers to simplify the processing of concatenated input documents, leverages the Longformer (Beltagy et al., 2020) to pre-train with a novel entity-based sentence masking objective. However, computing token-wise self-attention in the Transformer takes pairs of token relations into account but lacks syntactic support that may cause content irrelevance and deviation for summary generation (Jin et al., 2020).

Table 1: Generated summaries via different MDS models. Different colors mean different thought groups.

Source Documents	a girl reported missing more than two years ago when she was 15 told police she escaped a home in illinois ... ... they recovered the child and arrested a 24-year-old man ... ... she was 15 when she disappeared. she escaped from the home in washington park earlier this week and went to police ...
HT	... she was also taken into custody. ...
FT	... the girl , who was 15 when she escaped from a home in washington park earlier this week. ...
ParsingSum-HT (Ours)	... a 24-year-old man were arrested and taken into custody. ...
ParsingSum-FT (Ours)	... she was 15 when she disappeared from the home. ...

Many research works seek to incorporate linguistic knowledge to further improve the quality of summaries. Daniel et al. (Daniel et al., 2007) suggested that linguistic knowledge help improve the informativeness of summaries. Sho et al. (Sho et al., 2016) proposed an attention-based encoder-decoder model that adopts abstract meaning representation parser to capture structural syntactic and semantic information. The authors also pointed out that for natural language generation tasks in general, semantic information obtained from external parsers could help improve the performance of encoder-decoder based neural network model. Patrick et al. (Patrick et al., 2019) adopted named entities and entity coreferences for summarization problem. Jin et al. (Jin et al., 2020) enriched a graph encoder with semantic dependency graph to produce semantic-rich sentence representations. Song et al. (Song et al., 2020) presented a LSTM-based model to generate sentences and the parse trees simultaneously by combining a sequential and a tree-based decoder for abstractive summarization generation.

Dependency parsing, an important linguistic knowledge that retains the intra-sentence syntactic relations between words, has been adopted and shown promising results in a variety natural language processing task (Hiroyuki et al., 2019; Sun et al., 2019; Kai et al., 2020; Cao et al., 2021; Wu et al., 2017). The parsing information is usually formed as a tree structure that offers discriminate syntactic paths on arbitrary sentences for information propagation (Sun et al., 2019). The grammatical structure between the pair of words can be extracted from the dependency parser helping the model retain the syntactic structure. Therefore, in this work, we introduce a generic and flexible framework linguistic guided attention to incorporate dependency information into the Transformer based summarization models. We develop the proposed framework based on Flat Transformer (FT) and Hierarchical Transformer (HT), named ParsingSum-FT and ParsingSum-HT. Our proposed models can also be applied for both single and multiple document summarization.

Table 1 is an example to illustrate why dependency information helps improve the quality of summaries. The data source is from Multi-News dataset (Alexander et al., 2019). The HT model can not distinguish who was arrested: it should be *"a 24-year-old man"* rather than *"she"*. In contrast, ParsingSum-HT (our model) shows consistent content with source documents. The potential reason is that the dependency parsing captures the relation between *"arrested"* and *"man"*, which keeps the token relations for summaries generation. We also find the FT model mingles two events within two sentences. However, the source documents show two events: (1) the disappearance of the girl in Illinois was at her age of 15; (2) she escaped from her Washington Park home two years later. Comparatively, ParsingSum-FT (our model) retains correct information. This is due to, from the linguistic perspective, a sentence is a linguistic unit that has complete meaning (Halliday et al., 2014). Furthermore, dependency parsing focuses on intra-sentence relations that help summaries retain correct syntactic structure. The main contributions of this paper include:

- We propose a simple yet effective linguistic-guided attention mechanism to incorporate dependency relations with multi-head attention. The proposed linguistic-guided attention can be seamlessly incorporated into multiple mainstream Transformer-based summarization models to improve their performances.

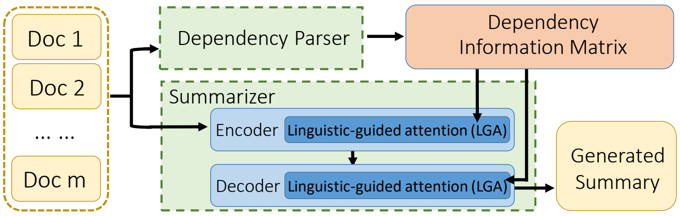

Doc 1
Doc 2
... ...
Doc m
Dependency Parser
Dependency Information Matrix
Summarizer
Encoder
Linguistic-guided attention (LGA)
Decoder
Linguistic-guided attention (LGA)
Generated Summary

Figure 1: The framework of ParsingSum. The set of documents are first fed into the encoder to generate the representations. In the meantime, these documents are input to a dependency parser to produce their sentence dependency information. The dependency information matrix will be further processed into a linguistic-guided attention mechanism and then fused with Transformer's multi-head attention to guide the downstream summary generation.

- We evaluate and compare the proposed model with several strong techniques. The results of automatic and human evaluation demonstrate that the models equipped with the linguistic-guided attention receive better performances over the compared models.

- We provide an extensive analysis of various settings and configurations of the proposed model. These results can help researchers understand the intuition of ParsingSum and serve as an informative reference for the summarization community.

2 Methodology

Figure 1 presents the framework of the proposed model ParsingSum. The proposed linguistic-guided attention mechanism is generic and flexible to be applied in different Transformer structures. Inside the model, the encoder is a representations learner to learn distinctive feature representations from the source documents and decoder is able to decipher representations into language domain for summary generation. More concretely, the document sets are first fed into a Transformer-based encoder for representation learning. Meanwhile, the source documents are passed into an external dependency parser to fetch the dependency relations. These relations and the Transformer's multi-head attention then be input into the linguistic-guided attention mechanism to construct the linguistic attention map. With the assistance of linguistic information, the model can grasp intra-sentence linguistic relations for summaries generation.

2.1 Dependency Information Matrix

Dependency grammar is a family of grammar formalisms that plays an important role in natural language processing. The dependency parser constructs several dependency trees that represent grammatical structure and the relations between *head* words and corresponding *dependent* words. To utilize these dependency information, we first adopt an external dependency parser (Dozat et al., 2017), which can handle sentences of any length, to generate a set of dependency trees from multiple documents. The trees contain dependencies between any pair of dependent words in one sentence. Let P denotes the dependency information matrix for one sentence. $p_{ij} \in P$ is a dependency weight between token t_i and token t_j. We simplify the definition of the weight as shown in Eq.(1):

$$p_{ij} = \begin{cases} 1 & t_i \ominus t_j \\ 0 & t_i \oslash t_j \end{cases} \quad (1)$$

where $t_i \ominus t_j$ indicates that t_i and t_j have a dependency relation, while $t_i \oslash t_j$ represents there is no dependency between the two tokens. To simplify the model, we consider the relations are undirected by ignoring the direction of *head* word and *dependent* word. For any pair of tokens, as long as there is a dependency between them, the dependency information matrix is assigned a value of 1, otherwise it will be set to 0. We hope to keep all dependency relations between the pair words in a simple yet effective manner.

2.2 Linguistic-Guided Attention Mechanism

In order to process source documents effectively and preserve salient source relations in the summaries, in

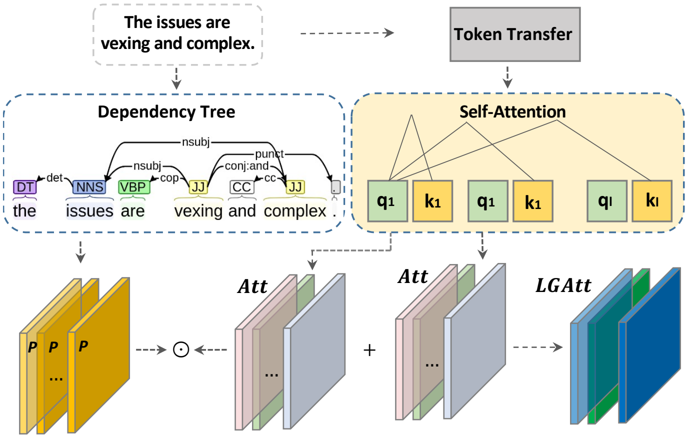


Figure 2: The linguistic-guided attention mechanism. The given exemplary sentence *The issues are vexing and complex.* is from Multi-News dataset (Alexander et al., 2019). Different properties of vocabularies and relations between words are included in the parsing information. The linguistic-guided attention mechanism incorporates the dependency information matrix P constructed from dependency trees of the input content and the Transformer's multi-head attention of this input content.

ParsingSum, we propose a novel linguistic-guided attention mechanism to extend the Transformer architecture (Ashish et al., 2017; Liu et al., 2019). Figure 2 depicts this mechanism on an exemplary sentence from Multi-News dataset (Alexander et al., 2019). linguistic-guided attention joins the dependency information matrix with the multi-head attention from source documents to generate syntactic-rich features. The linguistic-guided attention mechanism can be viewed as learning graph representations for the input sentences. Let $x_i^l \in \mathbb{R}^{d_{model*1}}$ denotes the output vector of the last encoding layer of Transformer for token t_i. For the attention head $head_z \in Head(j = 1, 2, ..., h)$, h represents the number of head. We have:

$$\begin{aligned} q_{i,head_z} &= W^{q,head_z} x_i^l \\ k_{i,head_z} &= W^{k,head_z} x_i^l \\ v_{i,head_z} &= W^{v,head_z} x_i^l \end{aligned} \tag{2}$$

where $W^{q,head_z}$, $W^{k,head_z}$, $W^{v,head_z} \in \mathbb{R}^{d_k * d_{model}}$ are weight matrices. d_k is the dimension of the key, query and value. $q_{i,head_z}, k_{i,head_z}, v_{i,head_z} \in \mathbb{R}^{d_k * 1}$ are sub-query, sub-key and sub-values in different heads and we concatenate them respectively:

$$\begin{aligned} Q_i &= concat(q_{i,head_1}, q_{i,head_2}, ..., q_{i,head_h}) \\ K_i &= concat(k_{i,head_1}, q_{i,head_2}, ..., q_{i,head_h}) \\ V_i &= concat(v_{i,head_1}, q_{i,head_2}, ..., q_{i,head_h}) \end{aligned} \tag{3}$$

where $Q_i, K_i, V_i \in \mathbb{R}^{h*d_k*1}$ are corresponding key, query and value for attention calculation. In ParsingSum, the linguistic-guided attention merges dependency information with multi-head attention in the following manner:

$$LGAtt_{ij} = \alpha M_{ij} \odot Att_{ij} + Att_{ij} \tag{4}$$

where

$$Att_{ij} = softmax\left(\frac{{Q_i}^T K_j}{\sqrt{d_k}}\right) \tag{5}$$

$$M_{ij} = Stack\_h(p_{ij}) \tag{6}$$

where α is a trade-off hyper-parameter to balance the linguistic-guided information M_{ij} and multi-head attention Att_{ij}. In order to fuse dependency weight p_{ij}, we build a function $stack\_h(\cdot)$ to repeat p_{ij} on the dimension of head to have the same size with $Att_{ij} \in \mathbb{R}^{h*1*1}$. $\odot$ denotes the element-wise Hadamard product. Then, we have:

$$Context_i = \sum_j LGAtt_{ij} \cdot V_j \tag{7}$$

where $Context_i$ represents the context vectors generated by linguistic-guide attention. Later on, two layer-normalization operations are applied to $Context_i$ to get the output vector of current encoder layer for token t_i:

$$x_i^{l+1} = LayerNorm\left(k_i + FFN\left(k_i\right)\right) \tag{8}$$

$$k_i = LayerNorm\left(x_i^l + Context_i\right) \tag{9}$$

where FFN is a two-layer feed-forward network with ReLU as activation function. Then, the learned feature representations are passed into multiple decoder layers that are fairly similar to the Flat Transformer structure (Sebastian et al., 2018).

3 Experiments

In this section, we report the effectiveness of the proposed linguistic-guided attention. Extensive analyses have been done on how to select suitable fusion weights in linguistic-guided attention, as well as the influence of batch size for model training. Later on, discussions on different fusion methods and their visualization are conducted.

3.1 Models for Comparison

We compare ParsingSum with the following models: *LexRank* computes the importance of a sentence-based on the concept of eigenvector centrality in a sentence graph (Gunes et al., 2004). *TextRank* is a graph-based ranking model (Rada et al., 2004). *Maximal Marginal Relevance (MMR)* (Jaime et al., 1998) considers the importance and redundancy of a sentence in a complementary way to decide whether to select the sentence for the summary. *SummPip* (Zhao et al., 2020) considers both linguistic knowledge and deep neural representations for summary generation. *BRNN*[1] is an bidirectional RNN-based model. *Flat Transformer (FT)* (Alexander et al., 2019) is a Transformer-based model on a flat token sequence. *Hi-MAP* (Alexander et al., 2019) incorporates MMR into a pointer-generator network. *Hierarchical Transformer (HT)* (Liu et al., 2019) is an abstractive summarizer that can capture cross-document relationships via hierarchical Transformer encoder and flat Transformer decoder.

3.2 Experimental Settings

We equip the proposed linguistic-guided attention on both Hierarchical Transformer (HT) and Flat Transformer (FT) architectures. Two models are thus derived: ParsingSum-HT and ParsingSum-FT. For ParsingSum-HT, we follow the implementation of the HT model by using six local Transformer layers and two global Transformer layers with eight heads[2]. For ParsingSum-FT, we follow FT model settings and adopt four encoder layers and four decoder layers[3]. For training, we use *Adam* optimizer ($\beta1$=0.9 and $\beta2$=0.998). The dropout rates of both encoder and decoder are set to 0.1. The initial learning rate is set to 1×10^{-3}. The first 8000 steps are trained for warming up and the models are trained with a multi-step learning rate reduction strategy. We evaluate the proposed model and compare its performances with multiple baseline models using ROUGE scores (Lin et al., 2004), the most commonly used evaluation metrics, and human evaluation. The experiments are conducted on two datasets : Multi-News dataset (Alexander et al., 2019) and WCEP-100 dataset (Demian et al., 2020). Multi-News a large-scale English MDS benchmark dataset

Table 2: Models comparison on Multi-News test set. We rerun all the compared models under the same environment. The best results for each column are in bold.

Models	ROUGE-F		
	1	2	L
LexRank	37.92	13.10	16.86
TextRank	39.02	14.54	18.33
MMR	42.12	13.19	18.41
SummPip	42.29	13.29	18.54
BRNN	38.36	13.55	19.33
FT	42.98	14.48	20.06
Hi-MAP	42.98	14.85	20.36
HT	36.09	12.64	20.10
ParsingSum-HT (Ours)	37.34	13.00	20.42
ParsingSum-FT (Ours)	**44.32**	**15.35**	**20.72**

Table 3: Models comparison on WCEP-100 test set. The best results for each column are in bold.

Models	ROUGE-F		
	1	2	L
HT	23.20	5.78	17.45
FT	23.41	6.64	17.93
ParsingSum-HT (Ours)	24.03	6.42	18.31
ParsingSum-FT (Ours)	**26.45**	**7.06**	**18.98**

[1] We implement the BRNN model based on https://github.com/Alex-Fabbri/Multi-News/tree/master/code/OpenNMT-py-baselines

[2] We train the HT model on one GPU for 100,000 steps with batch-size 13,000.

[3] We implement the FT model based on https://github.com/Alex-Fabbri/Multi-News/tree/master/code/OpenNMT-py-baselines. We train the FT model for 20,000 steps with batch-size 4096 on one GPU.

Table 4: The analysis of fusion weights of linguistic-guided attention on Multi-News validation set. The best results for each column are in bold.

Models	ROUGE-F		
	1	2	L
HT	36.02	12.57	20.05
ParsingSum-HT (α=1)	36.71	12.79	20.27
ParsingSum-HT (α=2)	35.64	12.18	19.80
ParsingSum-HT (α=3)	**36.74**	**12.86**	**20.29**
FT	42.81	14.25	19.81
ParsingSum-FT (α=1)	43.69	14.67	19.95
ParsingSum-FT (α=2)	**43.84**	**15.01**	**20.50**
ParsingSum-FT (α=3)	43.61	14.92	20.13

extracted from news articles. It includes 56,216 article-summary pairs and it is further scattered with the ratio 8:1:1 for training, validation and testing respectively. Each document set contains 2 to 10 articles with a total length of 2103.49 words. The average length of golden summaries is 263.66 words. WCEP-100 consists of 10,200 document sets (8158 for training, 1020 for validation and 1022 for testing) with one corresponding human-written summary. The average length of the summaries are 32 words. Deep Biaffine dependency parsing (Dozat et al., 2017) are used to generate dependency information for these source documents.

3.3 Overall Performance

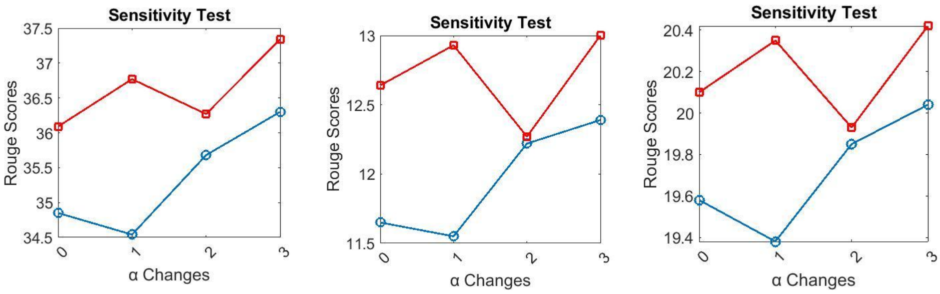
We evaluate the proposed ParsingSum-HT, ParsingSum-FT and compare them with multiple mainstream models on both Multi-News and WCEP-100 datasets. For fair comparisons, we rerun all the compared models under the same environment. For Multi-News dataset, as shown in Table 2, the ParsingSum-HT model receives higher ROUGE scores (across all ROUGE-1, ROUGE-2 and ROUGE-L) steadily compared to the original HT model. The linguistic-guided attention helps the model raise 1.25 on ROUGE-1 score, 0.36 on ROUGE-2 score, and 0.32 on ROUGE-L respectively. It indicates the outstanding capability of ParsingSum models to retain the intention of original documents when generating summaries. A similar phenomenon shows on the ParsingSum-FT model. More specifically, ParsingSum-FT surpasses FT model 1.34 on ROUGE-1 score, 0.87 on ROUGE-2 score, and 0.66 on ROUGE-L score, which shows the effectiveness of linguistic-guided attention on the Transformer-based models. It is worth noting that the proposed ParsingSum-FT is able to outperform its baseline (i.e., FT model) by a large margin and also receives the highest ROUGE scores across all the compared methods. The effect of linguistic-guided attention can be verified on the WCEP-100 dataset. The ROUGE results can be improved on both two version of Transformer based summarization models. These results indicate the outstanding capability of linguistic-guided attention to retain the intention of original documents when generating summaries.

Table 5: Human evaluation results on the Multi-News dataset. The best results for each column are in bold.

Models	Fluency	Informativeness	Consistency
Hi-MAP	2.53	2.80	2.33
FT	2.47	2.67	2.60
HT	2.20	2.13	2.40
ParsingSum-HT	2.73	**2.93**	**2.87**
ParsingSum-FT	**2.87**	2.87	2.73

3.4 Human evaluation

Although ROUGE are the standard evaluation metrics for summarization tasks, they focus on lexical matching instead of semantic matching. Therefore, in addition to the automatic evaluation, we access model performance by human evaluation in a semantic way. We invite three annotators who research natural language processing to evaluate the performance of five models (Hi-MAP, FT, HT, ParsingSum-FT, ParsingSum-HT) independently. For each model, 30 summaries are randomly selected from the Multi-News dataset. Three criteria are taken into account to evaluate the quality of generated summaries: (1) Informativeness: how much important information does the generated summary contain from the input document? (2) Fluency: how coherent are the generated summaries? (3) Consistency: how closely the information in the generated summaries are consistent with the input documents? Annotators are asked to give scores from 1 (worst) to 5 (best). Table 5 summarizes the comparison results of five summarization models. For each model, the score of each criterion is computed by averaging the score of all summary samples. The re-



(ROUGE-1) (ROUGE-2) (ROUGE-L)

Figure 3: The performance of ParsingSum-HT on small (in blue) and large batch-size setting (in red).

Table 6: Performance of ParsingSum-HT via different fusion methods on Multi-New validation set. The best results for each column are in bold.

Models	ROUGE-F		
	1	2	L
ParsingSum-HT (P0.25)	19.50	3.40	12.59
ParsingSum-HT (G0.25)	16.84	1.92	11.36
ParsingSum-HT (G8)	20.18	3.55	13.00
ParsingSum-HT (α=3)	**36.74**	**12.86**	**20.29**

sults demonstrate that the Transformer based models equipped with linguistic-guided attention are able able to generate higher quality summaries than the baseline models in terms of informativeness, fluency, and consistency. These human evaluation results further validate the effectiveness of our proposed linguistic-guided attention mechanism.

3.5 Analysis

We further analyze the effects of the trade-off parameter α and batch-size in ParsingSum. We also examine and discuss different manners to incorporate parsing information into the proposed model.

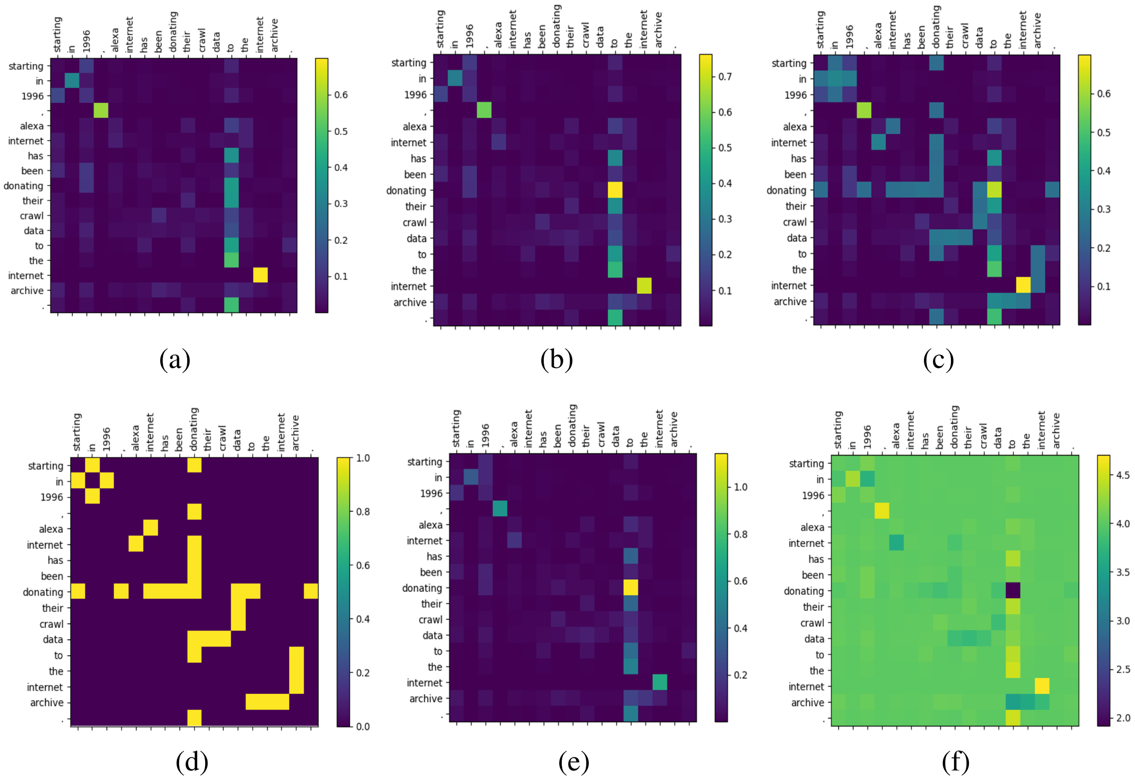
The Analysis of the Fusion Weights. The trade-off factor α controls the intensity of attention from a linguistic perspective to be fused with multi-head attention. To analyze its importance, we conduct experiments by setting α to 0, 1, 2, and 3 ($\alpha = 0$ denotes the naive Transformer model without linguistic-guided attention) on the two proposed models on the validation set. The results are shown in Table 4. Generally, there is an increasing trend with the increment of α. This rising trend further proves assigning a relatively larger α in a suitable range can improve the performance of summarization models.

The Analysis of Batch-size. Batch-size is considered to have a great effect on the mini-batch stochastic gradient descent process of model training (Smith et al., 2018) and it will thus further affect the model performance. To validate it empirically, we train the model with small/large batch-size (the small batch-size is 4,500 and the large one is 13,000) of the ParsingSum-HT model. The experiments are conducted with different α. The results in Figure 3 show that smaller batch-size reduces the performance on all the evaluation metrics. Interestingly, the ROUGE scores of the small batch-size setting are steadily increasing with α changes from 1 to 3; when the model is trained with large batch-size, the increasing trend is retained but the ROUGE scores are jittering when α equals two. It indicates different batch-sizes have different sensitivities towards the change of α.

The Analysis of the Fusion Methods. How to integrate the parsing information into the Transformer-based model is important in our work. In addition to the fusion method introduced in Section 2.2, we attempt several other fusion methods under a small batch-size setting of the ParsingSum-HT model: (1) Direct fusion. Weight the dependency parsing matrix and add it directly to the multi-head attention. It denotes as ParsingSum-HT (P0.25):

$$LGAtt_{ij} = 0.25M_{ij} + Att_{ij} \qquad (10)$$

(2) Gaussian-based fusion. We adopt the idea from (Li et al., 2020) and apply Gaussian weights to the product of the dependency information and the multi-head attention. The Gaussian weights are set to 0.25 (ParsingSum-HT (G0.25)) and 8



(a) (b) (c)

(d) (e) (f)

Figure 4: Visualization of different fusion methods. (a) HT model; (b) ParsingSum-HT (α=1); (c) ParsingSum-HT (P0.25); (d) dependency parsing matrix; (e) ParsingSum-HT (α=3); (f) ParsingSum-HT (G0.25).

(ParsingSum-HT (G8)):

$$LGAtt_{ij} = \frac{(1 - M_{ij}Att_{ij})^2}{0.25} + Att_{ij} \tag{11}$$

$$LGAtt_{ij} = \frac{(1 - M_{ij}Att_{ij})^2}{8} + Att_{ij} \tag{12}$$

Figure 4(a) and 4(d) represent the heatmap of the HT model and dependency parsing matrix. Figure 4(b), 4(c), 4(e), and 4(f) illustrate the attention maps of different fusion methods. Table 6 presents the performance of the mentioned fusion methods on Multi-New validation set. ParsingSum-HT with α=3 receives the best results for all ROUGE scores. The potential reason is that through direct fusion and Gaussian fusion, the scale of the original multi-head attention has been overwhelmed, leading to posing the dependency information in a dominant position. In this case, the normal gradient backpropagation process has been disturbed. The experiment results indicate that a direct summation of the weighted dependency parsing matrix and multi-head attention may damage the original attention. On the other hand, a "soft" fusion (when α is adopted) of these two attentions can achieve promising results.

4 Conclusion

This paper presents a generic framework to leverage linguistic knowledge to improve the performance of abstractive Transformer-based summarization models. The proposed linguistic guided attention mechanism can be seamlessly incorporated into multiple mainstream Transformer-based summarization models and can be outperform existing Transformer-based methods by a large margin. We develop two models based on Flat Transformer (FT) and Hierarchical Transformer (HT). The proposed ParsingSum-HT and ParsingSum-FT incorporate dependency relations with Transformer's multi-head attention for summaries generation. The experiments confirm that utilizing dependency information from the source documents is beneficial to guide the summaries generation process.

References

Perez-Beltrachini, L. & Lapata, M. Multi-Document Summarization with Determinantal Point Process Attention. *Journal Of Artificial Intelligence Research.* **71**, pp. 371-399 (2021)

Dhakras, P. & Shrivastava, M. BoWLer. A Neural Approach to Extractive Text Summarization. *Proceedings Of The 32nd Pacific Asia Conference On Language, Information And Computation* (2018)

Xiao, W., Beltagy, I., Carenini, G. & Cohan, A. PRIMERA: Pyramid-based Masked Sentence Pretraining for Multi-document Summarization. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (2022)

Beltagy, I., Peters, M. & Cohan, A. Longformer: The long-document Transformer. *ArXiv Preprint ArXiv:2004.05150* (2020)

Deguchi, H., Tamura, A. & Ninomiya, T. Dependency-Based Self-Attention for Transformer NMT. *Proceedings Of The International Conference On Recent Advances In Natural Language Processing*. pp. 239-246 (2019)

Ghalandari, D., Hokamp, C., Pham, N., Glover, J. & Ifrim, G. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 1302-1308 (2020)

Smith, S., Kindermans, P., Ying, C. & Le, Q. Don't Decay the Learning Rate, Increase the Batch Size. *Proceedings Of The 6th International Conference On Learning Representations* (2018)

Zhao, J., Liu, M., Gao, L., Jin, Y., Du, L., Zhao, H., Zhang, H. & Haffari, G. SummPip: Unsupervised Multi-Document Summarization with Sentence Graph Compression. *Proceedings Of The 43rd International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 1949-1952 (2020)

Halliday, M., Matthiessen, C., Halliday, M. & Matthiessen, C. An Introduction to Functional Grammar. *Routledge* (2014)

Gehrmann, S., Deng, Y. & Rush, A. Bottom-Up Abstractive Summarization. *Proceedings Of The 2018 Conference On Empirical Methods In Natural Language Processing*. pp. 4098-4109 (2018)

Wang, K., Shen, W., Yang, Y., Quan, X. & Wang, R. Relational Graph Attention Network for Aspect-based Sentiment Analysis. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 3229-3238 (2020)

Liu, P., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. & Shazeer, N. Generating Wikipedia by Summarizing Long Sequences. *Proceedings Of 6th International Conference On Learning Representations* (2018)

Cao, Q., Liang, X., Li, B. & Lin, L. Interpretable Visual Question Answering by Reasoning on Dependency Trees. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **43**, 887-901 (2021)

Wu, S., Zhang, D., Yang, N., Li, M. & Zhou, M. Sequence-to-Dependency Neural Machine Translation. *Proceedings Of The 55th Annual Meeting Of The Association For Computational Linguistics*. pp. 698-707 (2017)

Sun, K., Zhang, R., Mensah, S., Mao, Y. & Liu, X. Aspect-Level Sentiment Analysis Via Convolution over Dependency Tree. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing*. pp. 5678-5687 (2019)

Carbonell, J. & Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Proceedings Of The 21st Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 335-336 (1998)

Jin, H., Wang, T. & Wan, X. Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 6244-6254 (2020)

Fabbri, A., Li, I., She, T., Li, S. & Radev, D. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. *Proceedings Of The 57th Conference Of The Association For Computational Linguistics*. pp. 1074-1084 (2019)

Li, W., Xiao, X., Liu, J., Wu, H., Wang, H. & Du, J. Leveraging Graph to Improve Abstractive Multi-Document Summarization. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 6232-6243 (2020)

Fernandes, P., Allamanis, M. & Brockschmidt, M. Structured Neural Summarization. *Proceedings Of The 7th International Conference On Learning Representations* (2019)

Takase, S., Suzuki, J., Okazaki, N., Hirao, T. & Nagata, M. Neural Headline Generation on Abstract Meaning Representation. *Proceedings Of The 2016 Conference On Empirical Methods In Natural Language Processing*. pp. 1054-1059 (2016)

Song, K., Lebanoff, L., Guo, Q., Qiu, X., Xue, X., Li, C., Yu, D. & Liu, F. Joint Parsing and Generation for Abstractive Summarization. *Proceedings Of The Thirty-Fourth AAAI Conference On Artificial Intelligence*. pp. 8894-8901 (2020)

Lin, C. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings Of The Workshop Of Text Summarization Branches Out*. pp. 74-81 (2004)

Leite, D., Rino, L., Pardo, T. & Nunes, M. Extractive Automatic Summarization: Does more Linguistic Knowledge Make a Difference? *Proceedings Of The 2rd Workshop On TextGraphs: Graph-based Algorithms For Natural Language Processing*. pp. 17-24 (2007)

Liu, Y. & Lapata, M. Hierarchical Transformers for Multi-Document Summarization. *Proceedings Of The 57th Conference Of The Association For Computational Linguistics*. pp. 5070-5081 (2019)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention is All you Need. *Proceedings Of The Annual Conference On Neural Information Processing Systems*. pp. 5998-6008 (2017)

Mihalcea, R. & Tarau, P. TextRank: Bringing Order into Text. *Proceedings Of The 2004 Conference On Empirical Methods In Natural Language Processing*. pp. 404-411 (2004)

Erkan, G. & Radev, D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal Of Artificial Intelligence Research*. **22** pp. 457-479 (2004)

Ma, C., Zhang, W., Guo, M., Wang, H. & Sheng, Q. Multi-document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys* (2022)

Dozat, T. & Manning, C. Deep Biaffine Attention for Neural Dependency Parsing. *Proceedings Of The 5th International Conference On Learning Representations* (2017)

Jin, H., Wang, T. & Wan, X. SemSUM: Semantic Dependency Guided Neural Abstractive Summarization. *Proceedings Of The Thirty-Fourth AAAI Conference On Artificial Intelligence*. pp. 8026-8033 (2020)

VART: Vocabulary Adapted BERT Model for Multi-label Document Classification

Zhongguang Zheng
Fujitsu R&D Center
zhengzhg@fujitsu.com

Lu Fang
Fujitsu R&D Center
fanglu@fujitsu.com

Yiling Cao
Fujitsu R&D Center
caoyiling@fujitsu.com

Jun Sun
Fujitsu R&D Center
sunjun@fujitsu.com

Abstract

Large scale pre-trained language models (PTLMs) such as BERT have been widely used in various natural language processing (NLP) tasks, since PTLMs greatly improve the downstream task performances by fine-tuning the parameters on the target task datasets. However, in many NLP tasks, such as document classification, the task datasets often contain numerous domain specific words which are not included in the vocabulary of the original PTLM. Those out-of-vocabulary (OOV) words tend to carry useful domain knowledge for the downstream tasks. The domain gap caused by OOV words may limit the effectiveness of PTLM. In this paper, we present VART, a concise pre-training method to adapt BERT model by learn OOV word representations for multi-label document classification (MLDC) task. VART employs an extended embedding layer to learn the OOV word representations. The extended layer can be pre-trained on the task datasets with high efficiency and low computational resource. The experiments for MLDC task on three datasets from different domains with different sizes demonstrate that VART consistently outperforms the conventional PTLM adaptation methods such as fine-tuning, task adaption and other pre-trained model adaptation methods.

1 Introduction

Pre-trained language models (PTLMs) such as GPT (Radford and Narasimhan, 2018) and BERT (Devlin et al., 2018), which are trained on massive unlabeled datasets, can effectively encode rich knowledge into huge parameter spaces. Therefore, by fine-tuning the PTLM parameters, the encoded knowledge is able to benefit a wide range of downstream natural language processing (NLP) tasks (Dai et al., 2021; Liang et al., 2020; Adhikari et al., 2019; Wang et al., 2019; Zhu et al., 2020; Gao et al., 2019).

However, applying PTLMs to the specific domain tasks always faces the domain gap problem. Conventionally, PTLMs are trained on a large volume of general domain datasets with a fixed vocabulary extracted from the datasets. When applying such general domain PTLMs on a specialized domain dataset, e.g., patent documents, the domain gap becomes an important factor that hinders the performance of PTLMs. One of the causes of the domain gap is the domain words which are not included in the PLTM vocabulary. Although a PTLM is capable to handle the out-of-vocabulary (OOV) words by splitting each OOV word into multiple in-vocabulary sub-words, for instance, the word "chalcogenide" commonly seen in the patent document will be split into five sub-words including "ch", "##al", "##co", "##gen" and "##ide" with the vocabulary of BERT-base-cased model. As a result, the representation of "chalcogenide" is divided into five embedded vectors. Consequently, the information of "chalcogenide" which is intuitively preferable as an integral representation would be ignored in the downstream tasks. Since those OOV words tend to carry rich domain knowledge, the information loss potentially limits the effectiveness of PTLMs for tasks such as multi-label document classification (MLDC), which is known as a fundamental and essential NLP task and has been widely applied in specific domain tasks such as clinical code prediction (Scheurwegs et al., 2017; Mullenbach et al., 2018) for electronic health record (EHR) texts and biomedical document classification (Du et al., 2019; Baker and Korhonen, 2017). The datasets for domain-specific MLDC tasks always contain a large number of domain OOV words which challenge the use of PTLMs.

In order to bridge the domain gap for PTLMs, a plenty of researches have been conducted. Some of the prior researches have focused on fine-tuning PTLMs for text classification (Sun et al., 2019), while other researches have addressed the issue by adapting the PTLMs to the target domain datasets by training PTLMs from scratch with a new vocabulary, such as SciBERT (Beltagy et al., 2019). However, the fine-tuning approach is focused on adapting the PTLM parameters to the target domain and leaves aside the OOV issue, while training PTLMs from scratch sets an extremely high demand of computational resource and it is time-consuming. To solve the domain OOV problem efficiently, there are recent researches focusing on extending the original PTLM vocabulary with target domain words, such as exBERT (Tai et al., 2020), which complements the original BERT model with another BERT model for learning the OOV representations. However, exBERT model still faces the problems of efficiency and training complexity.

Inspired by the work of exBERT, we propose VART, a **V**ocabulary **A**dapted BE**RT** model which adapts the original BERT model by extending the vocabulary with the target domain vocabulary. Specifically, we just extend the embedding layer of BERT model to learn the OOV word representations while inheriting other BERT layers and then pre-train the model on the downstream task datasets. Comparing with exBERT, the main contributions of this work are summarized as follows:

- We demonstrate a concise training method to extend the BERT vocabulary with domain OOV words. VART overcomes the problems that remained in exBERT by boosting the efficiency in both pre-training and fine-tuning phases while saving computational resources.
- Extensive experiments are conducted on three datasets with different sizes and domains for MLDC task. Although smaller in model size, VART consistently outperforms exBERT and other baseline methods even on an extremely small scale task dataset.

2 Related Work

Fine-tuning PTLMs. The most common conventional approach for utilizing PTLMs is fine-tuning. Generally, fine-tuning is performed by replacing the output layer of a PTLM with other layers which are specified according to the downstream tasks. The parameters of the original PTLM are preserved and tuned on the task datasets. Various fine-tuning methods of BERT specially on document classification task are investigated in (Sun et al., 2019), such as studying the effectiveness of different BERT layers in the fine-tuning phase. Besides, a multi-task learning mechanism is also used to fine-tune the BERT model. Rietzler et al. (2019) fine-tunes the BERT model for sentiment classification task. Different adaptation scenarios such as in-domain, cross-domain and joint-domain are studied in the experiments. Gururangan et al. (2020) focuses on dataset selection for further pre-training the RoBERTa (Liu et al., 2019) model. Various dataset selection strategies, such as domain dataset selection (domain-adaptive pre-training, DAPT) and task dataset selection (task-adaptive pre-training, TAPT) are proposed. The experiment results demonstrate that an adapted PTLM is beneficial to various downstream NLP tasks.

Domain specific PTLMs. In order to further improve the performances on domain specific tasks, such as biomedical domain, researches focusing on training domain specific PTLMs have been proposed in recent years. SCIBert (Beltagy et al., 2019) leverages pre-training on large multi-domain scientific publications with a new in-domain vocabulary. The experiment shows that SCIBert outperforms general domain BERT in the scientific domain tasks. Moreover, the in-domain vocabulary is proven helpful in the experiment. Gu et al. (2020) pre-trains the domain BERT model from scratch with a customized vocabulary on the PubMed articles. BioBERT (Lee et al., 2019) inherits the vocabulary and the model

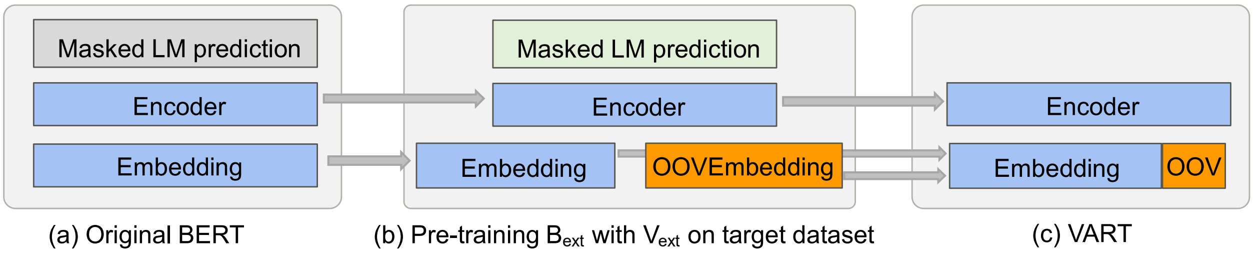


Figure 1: Overview of adapting BERT model by extending the vocabulary. We will reuse the original BERT encoder layer and further pre-train it with an extended vocabulary V_{ext} from the target dataset. While in the pre-training phase, we train the OOVEmbedding layer and MLM Prediction layer from scratch. Finally, the VART model is constructed by merging the two embedding layers into one.

parameters from original BERT model and then is further pre-trained on a large volume dataset that mainly consisted of PubMed articles. Similarly, Lee and Hsiang (2020) obtains a BERT model in patent domain by fine-tuning original BERT model on over two million patent documents.

Extending PTLM vocabulary. It is known that training PTLMs from scratch requires powerful computational resources and is time-consuming. Recently proposed exBERT (Tai et al., 2020) introduces a general training method to extend the original BERT from the general domain to a specific domain. The exBERT model preserves the original BERT model and vocabulary. Meanwhile, another smaller (or full size) BERT model is used to learn the information of domain OOV words. However, there are mainly two problems of this dual BERT model structure. Firstly, exBERT is much larger than a single BERT model so that it is highly inefficient in pre-training phase. For alleviating the problem, the author proposes tradeoffs such as shrinking the size of the extra BERT model and fixing the parameters of the original BERT model in the pre-training phase. However, the smaller size BERT model may be inadequate to learn document representations, while the original BERT model still faces the domain gap if the parameters are fixed. Secondly, it is non-trivial to combine the outputs from the two BERT encoders. For this reason, a weighting block comprised of a fully-connected layer and sigmoid activator is used to combine the two encoder outputs. This increases the training complexity.

The fine-tuning PTLMs method focuses on the model parameter adaptation and leaves aside the OOV issue. Training domain specific PTLMs from scratch is computational expensive and time-consuming. Inspired by exBERT, we propose VART to adapt the original BERT model to the target domain by learning the representations of domain OOV words. In order to solve the remained problems in exBERT, we use only one single BERT model with a minor modification for the adaptive pre-training. Comparing with exBERT, our model boosts the efficiency in both pre-training and fine-tuning phases without sacrificing the performances for MLDC task.

3 Methodology

In this section, we introduce VART model which is pre-trained with an extra OOV list from the target dataset illustrated in Figure 1. In the first place, we would like to define the PTLM adaptation task in this paper. Given an original pre-trained BERT model B with a vocabulary V_{in} and a training dataset D_t for MLDC in a specific domain, we expand V_{in} to V_{ext} with an OOV list extracted from D_t at first, and then design an extended BERT model B_{ext}, which is inherited from B and is further pre-trained with V_{ext}. Finally, VART is derived from B_{ext} and is fine-tuned for downstream MLDC task. In the rest of this paper, the term “BERT” refers to “BERT-base-cased”[1] model from Huggingface[2] repository and is implemented by Huggingface transformers[3] (Wolf et al., 2020).

[1] https://huggingface.co/bert-base-cased

[2] https://huggingface.co/

[3] https://github.com/huggingface/transformers

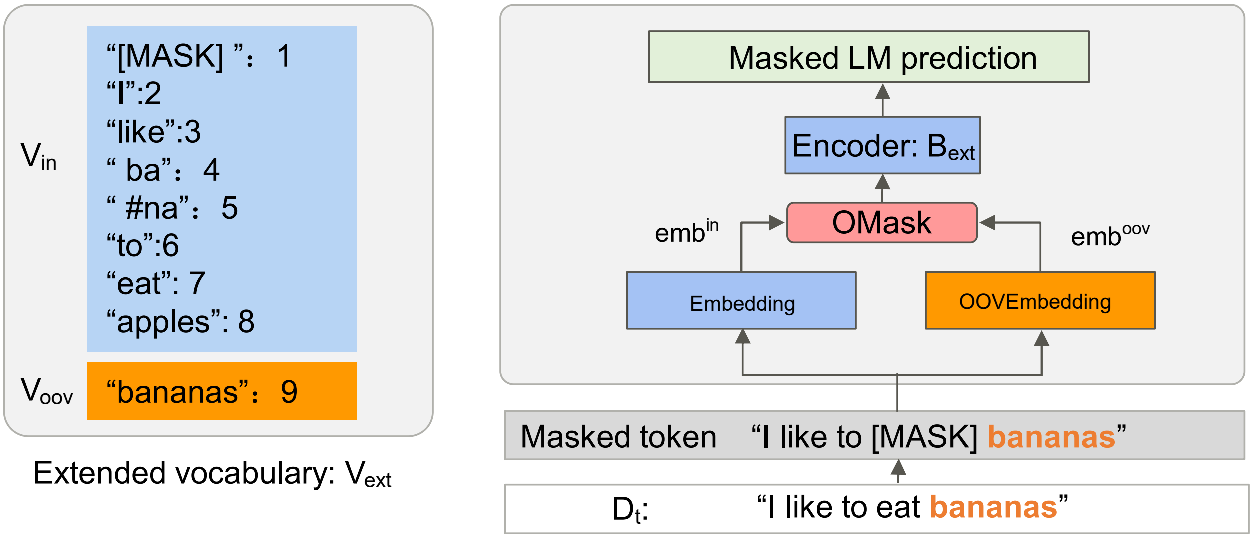


Figure 2: Overview of further pre-training B_{ext} with extended vocabulary and target dataset D_t. The $OMask$ (OOV Mask) selects and combines embeddings emb^{in} from the original Embedding layer for the existing vocabularies and emb^{oov} from OOVEmbedding layer for the new vocabularies.

OOV Extraction. Given a training dataset D_t of a specific task (MLDC for this paper), we first extract a word list V_t from D_t via WordPiece (Wu et al., 2016) algorithm, then an OOV vocabulary V_{oov} is constructed by selecting all the terms in V_t but not in V_{in}. Afterwards, V_{ext} is obtained by appending V_{oov} to V_{in}. As the example shown in Figure 2, we extract a new word "bananas" from D_t and append the word to V_{in}.

Pre-training VART. The structure of original BERT model can be categorized mainly into three components shown in Figure 1(a). The embedding layer encodes the input tokens into a real-valued vector. The encoder is a stack of Transformer (Vaswani et al., 2017) blocks consist of multiple self-attention heads and learns the final representation of the input sequence. The task layer generates the output for self-supervised tasks such as masked language model (MLM) and next sentence prediction (NSP).

In order to learn the representations for the OOV words, we simply complement an extra embedding layer to the original BERT model B while preserving the original embedding layer to encode the in-vocabulary words. As the example shown in Figure 2, by applying $OMask$, only the new word "bananas" is encoded by the OOVEmbedding layer. Afterwards, it is handy to use $OMask$ for combining the embedded vectors emb^{in} and emb^{oov}. This process is described in the following Equations, where w_i is id of the i-th word in the input sequence.

$$\mathrm{emb}_i^{\mathrm{in}} = \mathrm{Embedding}((1 - \mathrm{OMask}_i) \cdot w_i) \quad (1)$$

$$\mathrm{emb}_i^{\mathrm{oov}} = \mathrm{OOVEmbedding}(\mathrm{OMask}_i \cdot w_i) \quad (2)$$

$$\mathrm{emb}_i = (1 - \mathrm{OMask}_i)\mathrm{emb}_i^{\mathrm{in}} + \mathrm{OMask}_i\mathrm{emb}_i^{\mathrm{oov}} \quad (3)$$

The encoder of B_{ext} is initialized with the encoder in B. For the task layer, we only select MLM to pre-train B_{ext} in this work. Considering that the vocabulary size of V_{ext} is enlarged, the magnitude of prediction vector from the task layer should be increased to match the size of V_{ext}. Therefore, we create a new task layer with the proper dimension for pre-training B_{ext} on D_t. Considering that the extra embedding layer and the task layer in B_{ext} need to be trained from scratch, while the encoder layer is inherited from B, it is plausible to set different learning rates for those layers in the pre-training process. Afterwards VART model is obtained by concatenating the two embedding layers after pre-training B_{ext}(see Figure1(c)). At this point, we adapt B to the target dataset D_t by extending the vocabulary.

The final transformation brings another merit of VART in that our model is converted into a standard

Table 1: Datasets overview. The column "Len" denotes the average document length of each dataset. "Labels" is the actual label number of the dataset.

Dataset	Train	Test	Len	Labels	Domain
CEI	2.4k	1.2k	277	103	Clinical
EU-Leg	45k	6k	538	4193	Law
Patent	90k	10k	66	9152	Patent

BERT model from the non-standard structured B_{ext}. A standard structured BERT model is much more friendly to people who would deploy VART in their own applications since they do not need to import any extra libraries in their project reducing the risk of library conflicts.

4 Experiment

4.1 Datasets

We conduct experiments on three datasets for the MLDC task with different sizes and domains. An overview of all the datasets is shown in Table 1.

CEI dataset[4] (Larsson et al., 2017) is annotated for chemical exposure assessments[5]. The dataset contains 3.7k abstracts from PubMed documents and is categorized by experts into 32 classes denoting chemical exposure information.

EU-Leg dataset (Chalkidis et al., 2019) comprises 57k legislative documents from EURLEX[6]. The documents are annotated with multiple concepts from EUROVOC[7] and contain about 4.3k labels in total.

Patent dataset[8] (Huang et al., 2019) is collected from USPTO[9] containing 100k patent documents, including titles and abstracts. The hierarchical annotation category contains almost 9k labels.

4.2 Implementation

Pre-training. We use the exBERT library[10] and modify the training script to support the VART model. The BERT-base-cased model, which contains 12 Transformer layers with 12 self-attention heads and 768 hidden dimension, is selected as the original BERT model. The maximum input sequence length is set to 512 and only the masked language model task is chosen to pre-train all the models.

Table 2: Learning rate settings for pre-training different models. "BERT$_{tapt}$" denotes the task adapted training for BERT model. The subscript of exBERT and VART models denotes the dataset on which the learning rate is applied.

Models	Original Layers	Extended Layers
BERT$_{tapt}$	4e-05	–
exBERT$_{other}$	2e-05	1e-04
exBERT$_{CEI}$	4e-05	1e-04
VART$_{other}$	4e-05	1e-04
VART$_{EU\_Leg}$	2e-05	2e-04

For the extended encoder in exBERT model, we inherit the best settings in the author's work with hidden size 252 and feed-forward layer size 1024 (about 33% of the original BERT model size). Different learning rates are set for the original BERT layers and the extended layers in both exBERT and VART models. The detailed learning rate settings are listed in Table 2. We pre-train all the models for 50 epochs on the CEI dataset, 40 epochs on the Patent dataset and 10 epochs on the EU-Leg dataset in considering of training efficiency. The models saved after the final epoch will be used for the MLDC task. The batch size is set to 4 on all the datasets. The Adam (Kingma and Ba, 2015) optimizer is applied to tune the parameters. All the experiments hereafter are conducted on our in-house servers with GeForce GTX 1080 Ti/2080 Ti GPUs.

Fine-tuning. In the fine-tuning process, the hidden state h of the first token [CLS] is considered as the document representation and is followed by a fully connected layer (task layer) which predicts the final labels. Different learning rates are set to the pre-trained models and the task layer respectively. 2e-05 is set to all the pre-trained models, while 2e-04 is set to the task layer on CEI dataset and 1e-04 is set on Patent and EU-Leg datasets. Moreover, a

[4] https://figshare.com/articles/dataset/Corpus_and_Software/4668229

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5336247/

[6] https://eur-lex.europa.eu/

[7] https://publications.europa.eu/en/web/eu-vocabularies

[8] https://drive.google.com/open?id=1So3unr5p_vlYq31gE0Ly07Z2XTvD5QlM

[9] https://www.uspto.gov/

[10] https://github.com/cgmhaicenter/exBERT

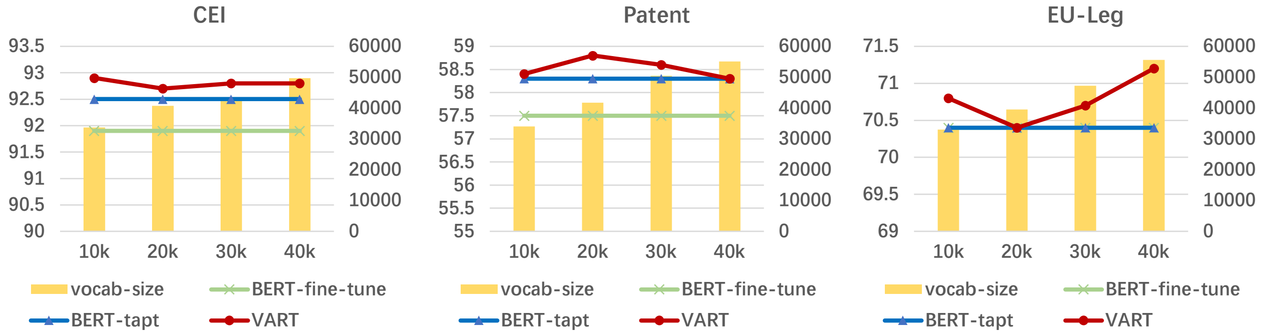


Figure 3: Visualization of results from different vocabularies. The line chart denotes the micro-F1 score corresponding to the left vertical axis and the histogram shows the vocabulary size corresponding to the right vertical axis.

Table 3: Overall experiment results. "Vocab Size" denotes the vocabulary size for training exBERT and VART models. For BERT models, the vocabulary size is 28,996.

Model	CEI	EU-Leg	Patent
BiGRU-LWAN	–	69.8%	–
HARNN	–	–	54.1%
BERT$_{fine\_tune}$	91.9%	70.4%	57.5%
BERT$_{tapt}$	92.6%	70.4%	58.3%
exBERT$_{ext}$	92.1%	**71.5%**	58.4%
exBERT$_{whole}$	92.8 %	70.8%	58.6%
VART	**92.9%**	71.2%	**58.8%**
Vocab Size	33,710	55,558	41,718

learning rate decay mechanism is adopted to boost the performance in the form of Equation 4, where α_0 is the initial learning rate and $decay\_rate$ is set to 0.9. We run 20 epochs for fine-tuning on all the datasets and the binary cross-entropy loss is used to train the classifier. Micro-F1 score is adopted as the evaluation metric.

$$\alpha' = \frac{1}{1 + decay\_rate \times epoch\_num}\alpha_0 \quad (4)$$

4.3 Baseline Methods

Networks without pre-trained models. HARNN (Huang et al., 2019) is a hierarchical attention-based RNN model. BiGRU-LWAN (Chalkidis et al., 2019) adopts a label-wise attention mechanism on the basis of a Bi-GRU layer. The two models carried out experiments on the same datasets as in this work and without pre-trained models. We cite the reported results on Patent and EU-Leg datasets from the works directly.

Table 4: Comparison of different vocabularies. Scores in bold denote the results better than Bert$_{tapt}$. The best scores are marked with †. The number in brackets presents the vocabulary size.

Settings	CEI	EU-Leg	Patent
Bert$_{fine-tune}$	91.9% (28,996)	70.4% (28,996)	57.5% (28,996)
Bert$_{tapt}$	92.6% (28,996)	70.4% (28,996)	58.3% (28,996)
VART$_{10k}$	**92.9**†**%** (33,710)	**70.8%** (33,000)	**58.4%** (34,008)
VART$_{20k}$	**92.7%** (40,726)	70.4% (39,507)	**58.8%**† (41,718)
VART$_{30k}$	**92.8%** (42,659)	**70.7%** (47,254)	**58.6%** (50,456)
VART$_{40k}$	**92.8%** (49,673)	**71.2%**† (55,558)	58.3% (55,075)

Fine-tune Only. The original BERT model is directly fine-tuned for the downstream MLDC task.

Task Adaptation. We follow the task adaptation (TAPT) method described in (Gururangan et al., 2020) to pre-train the BERT model on each training dataset with the vocabulary unchanged at first, and then fine-tune the adapted BERT model for the MLDC task.

exBERT. We inherit the best settings for the extended encoder in the author's work with hidden size 252, feed-forward layer size 1024, 12 attention heads and 12 hidden layers (about 33% of the orig-

Table 5: Efficiency and computational resource cost comparison of the pre-trained models.

		BERT$_{tapt}$	**exBERT**	**VART**
Parameter size		110M	147M	122M
FLOPs$_{pre\_train}$	CEI	55.2B	75B	60.6B
	Patent		74.6B	60.2B
	EU-Leg		80.1B	65.6B
FLOPs$_{fine\_tune}$	CEI Patent EU-Leg	43.5B	57.9B	43.5B
GPU usage		1	2	1

inal BERT model size). The exBERT model is pre-trained based on the same vocabulary with VART model. Moreover, there are two pre-training modes for exBERT: training the extended model only and training the whole model. We test the both pre-training modes in this work.

4.4 Experiment Results and Analysis

The overall experiment results are listed in Table 3. From the results, we can observe that the fine-tuning method greatly improves the performances with respect to the networks without BERT model. This confirms the effectiveness of BERT model, not surprisingly. Task adapted BERT model further improves the performance of fine-tuning method, which demonstrates that the adaptation of PTLMs is essential to improve the performance for specific tasks. VART achieves best scores on CEI and Patent datasets. Although exBERT produces the best result on EU-Leg dataset with training extended encoder mode, the result from VART on EU-leg is close to exBERT$_{ext}$ and is better that other baselines including exBERT$_{whole}$.

From the results we can also see the problems of exBERT. Firstly, the two training modes perform differently on different datasets. Moreover, the results of the two modes are also significantly different. On the CEI dataset, exBERT$_{whole}$ is significantly better than exBERT$_{ext}$, while the result is opposite on EU-Leg dataset. This indicates that it is difficult for exBERT to balancing the performances of the original BERT encoder and the extended encoder. On the contrary, VART model can be easily trained as a result of its simple structure.

Secondly, the training efficiency hinders the utilization of exBERT. Table 5 lists the comparison of the training efficiency and the computational resource cost between exBERT and VART. We can see that, although exBERT yields equivalent results on CEI and Patent datasets and produces the best result on EU-Leg dataset, VART is more efficient than exBERT in both pre-training and fine-tuning. For instance, as to the real running time, it took about 22 hours with VART on the Patent dataset comparing to 36 hours with exBERT$_{ext}$ and 40 hours with exBERT$_{whole}$. In the fine-tuning process, it took about 19 hours with VART model on EU-Leg dataset comparing to 40 hours with exBERT. Besides, VART has about 17% fewer parameters than exBERT without sacrificing performances. As to the computational resource cost, VART only requires 1 GPU for both pre-training and fine-tuning, while exBERT needs 2 GPUs under the same setting. All the evidences demonstrate that VART is able to further improve the performance on basis of conventional PTLM adaptation methods with a high efficiency and a low computational resource cost.

4.5 Impact of Vocabulary Size

We further conduct experiments to test VART model with different vocabulary sizes. API `BertWordPieceTokenizer` is used for extracting vocabularies from the target datasets. By setting the parameter `vocab_size` with different values, we can control the extracted vocabulary size. For each training dataset, we extract four vocabularies by setting `vocab_size` with 10k, 20k, 30k and 40k separately. The settings for pre-training and fine-tuning with different vocabularies remain the same as Section 4.2.

The detailed results of different vocabulary sizes are listed in Table 4, which is also visualized in Figure 3. From the results we can observe that VART model produces better results than BERT$_{tapt}$ model in most cases. This indicates that extending the vocabulary is beneficial to the MLDC task.

On the other hand, we could arrive at a similar conclusion with Tai et al. (2020) that increasing the vocabulary size may not always produce better results. The best score coming from the largest vocabulary can be only seen on the EU-Leg dataset. On the CEI and Patent datasets, the best scores are achieved with 10k and 20k vocabularies instead of using their largest vocabularies.

We hypothesis that the vocabulary size and the dataset size should to be proportional. Since larger vocabularies from smaller datasets may contain more low frequency words which provide less information for the MLDC task. However, the relationship between the vocabulary size and the training sample number is worth studying in the further.

5 Conclusion

We introduced VART, a concise pre-training method to extend the BERT model with domain OOV words. With a minor modification of adding an extra embedding layer to the original BERT model, we can adapt the BERT model to the target task datasets. Comparing with the conventional method such as exBERT, our approach maximizes the use of general domain BERT model with much higher efficiency, such as less pre-training time and lower computational resource requirements. Experiment results indicate that our approach leverages the domain gap of PTLMs for MLDC tasks. Since our approach is a general solution for adapting the BERT model, in the future we would like to examine VART in other NLP tasks.

References

Alec Radford and Karthik Narasimhan. 2018. *Improving language understanding by generative pretraining*. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/abs/1810.04805

Zhihao Dai, Zhong Li, and Lianghao Han. 2021. *BoneBert: A BERT-based Automated Information Extraction System of Radiology Reports for Bone Fracture Detection and Diagnosis. Advances in Intelligent Data Analysis XIX*, pp. 263–274. https://doi.org/10.1007/978-3-030-74251-5_21

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. *BoneBert: BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp 1054—1064.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. *DocBERT: BERT for Document Classification*. https://arxiv.org/abs/1904.08398

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. *Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering*. https://arxiv.org/abs/1908.08167

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. *Incorporating BERT into Neural Machine Translation*. https://arxiv.org/abs/2002.06823

Gao Zhengjie, Feng Ao, Song Xinyu, and Wu Xi. 2019. *Target-Dependent Sentiment Classification With BERT*. *IEEE Access*, vol. 7, pp. 154290–154299. 10.1109/ACCESS.2019.2946594

Elyne Scheurwegs, Boris Culea, Kim Luyckx, Léon Luyten, and Walter Daelemans. 2017. *Selecting relevant features from the electronic health record for clinical code prediction*. Journal of Biomedical Informatics, vol. 74, pp. 92–103. https://doi.org/10.1016/j.jbi.2017.09.004

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. *Explainable Prediction of Medical Codes from Clinical Text*. https://arxiv.org/abs/1802.05695

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. *ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association*, vol. 26, pp. 1279—1285. https://doi.org/10.1093/jamia/ocz085

Simon Baker and Anna Korhonen. 2017. *Initializing neural networks for hierarchical multi-label text classification*. BioNLP, https://doi.org/10.17863/CAM.12418

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. *How to Fine-Tune BERT for Text Classification?*. China National Conference on Chinese Computational Linguistics, pp 194–206. https://arxiv.org/abs/1905.05583

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. *Scibert: Pretrained contextualized embeddings for scientific text.* https://arxiv.org/abs/1903.10676

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. *exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. Findings of the Association for Computational Linguistics: EMNLP 2020*, pp 1433–1439. https://aclanthology.org/2020.findings-emnlp.129

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. *Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification.* https://arxiv.org/abs/1908.11860

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp 8342–8360. https://arxiv.org/abs/2004.10964

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A robustly optimized BERT pretraining approach.* https://arxiv.org/abs/1907.11692

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. *Domain specific language model pretraining for biomedical natural language processing.* https://arxiv.org/abs/2007.15779

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. *Biobert: pre-trained biomedical language representation model for biomedical text mining.* https://arxiv.org/abs/1901.08746

Jieh-Sheng Lee and Jieh Hsiang. 2020. *Patent classification by fine-tuning BERT language model.* https://doi.org/10.1016/j.wpi.2020.101965

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. *Google's neural machine translation system: Bridging the gap between human and machine translation.* https://arxiv.org/abs/1609.08144

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. *HuggingFace's Transformers: State-of-the-art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need. Advances in Neural Information Processing Systems*, pp 5998-–6008.

Kristin Larsson, Simon Baker, Ilona Silins, Yufan Guo, Ulla Stenius, Anna Korhonen, and Marika Berglund. 2017. *Text mining for improved exposure assessment. PLoS One.*

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Ion Androutsopoulos . 2019. *Large-Scale Multi-Label Text Classification on EU Legislation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp 6314–6322.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. *Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach. Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp 1051—1060.

Diederik P. Kingma and Jimmy Ba. 2015. *Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, Conference Track Proceedings.*

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. *Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach. Proceedings of the 28th ACM International Conference on Information, Knowledge Management*, pp 1051–1060.

Are Emoji, Sentiment, and Emotion Friends? A Multi-task Learning for Emoji, Sentiment, and Emotion Analysis

Gopendra Vikram Singh[*], Dushyant Singh Chauhan, Mauajama Firdaus[+], Asif Ekbal, and Pushpak Bhattacharyya[∓]
Department of Computer Science and Engineering
[*] Indian Institute of Technology Patna, India
{gopendra_1921cs15,1821cs17,asif}@iitp.ac.in
[∓] Indian Institute of Technology Bombay, India
pb@cse.iitb.ac.in

Abstract

Related tasks often have inter-dependence on each other and perform better when solved in a joint framework. For e.g., emoji ❤️ can help in the prediction of joy (happy) emotion and positive sentiment while 😡 can help in the prediction of angry, sad emotion and negative sentiment and so on. In this paper, we investigate the relationship between emojis, sentiment, and emotion by developing a multitask neural framework that performs emoji prediction (primary task) with the help of sentiment and emotion and their intensities (the auxiliary tasks). For our task at hand, we use the already available dataset (Emoji Analysis task @ SemEval 2018) which contains, along with tweets, emojis (😜😘😊) that conveys positive sentiment in general. We create an enriched version of this dataset named as SEEmoji (Sentiment and Emotion aware Emoji dataset) by collecting tweets, diverse emojis and labeling with the different kinds of sentiment and emotion classes. Empirical results on the SEEmoji dataset demonstrate that the proposed multitask framework yields better performance over the single-task learning.

1 Introduction

Humans are driven by emotions and, in everyday life, emotional outburst can be seen in various forms. With the popularity and growth of social media, people have access to the numerous platforms to voice their views, give opinions, and also express their feelings. With the advancement in artificial intelligence (AI), social media platforms such as Twitter, Facebook, Instagram etc. have brought people closer and, simultaneously, provided an opportunity to express their emotions in the best possible way. Presently, the number of users on social media worldwide is *3.81 billion* [1] and this number is increasing day-by-day. In addition, in recent times, social media users' writing patterns have also changed. They increased the use of pictographs, called emojis, along with the text, to make the message descriptive and lively.

No.	Utterances	Emoji	Sent	Emotion
1	*LoL @ West Covina, California*	😂	Pos	Joy
2	*Momma @ Disney's Magic Kingdom*	✨	Pos	Joy
3	*sooo sick of the snow ughh*	😡	Neg	Anger
4	*People make me sick*	🤮	Neg	Disgust
5	*Some are just so selfish*	🤮	Neg	Disgust

Table 1: Example to show the relationship between emoji, sentiment, and emotion.

Emoji is an essential aspect of daily conversation and adds more sense to language. Emoji is often used to convey thinly veiled disapproval humorously. This can be easily depicted through the example - *"Some are just so selfish* 🤮.". This tweet, at an outer glance, conveys that the person is extremely sad with some people's behaviour. But careful observation of the sentiment and emotion of the person helps us understand that the person is disgusted with these type of selfish people and has a negative sentiment during the tweet (c.f. 5^{th} tweet in Table 1).

Similarly, in this tweet, *"Momma @ Disney's Magic Kingdom* ✨*"*, the girl is extremely pleased after coming @ Disney's Magic Kingdom and careful observation of the sentiment and emotion of the girl helps us understand that the girl conveys joy (happy) emotion and positive sentiment in the tweet (c.f. 2^{nd} tweet in Table 1). This is where sentiment and emotion come into the picture.

In this paper, we exploit these relationships to make use of sentiment and emotion of the tweet for predicting emoji in a multi-task manner. The main contributions and/or attributes of our proposed research are as follows: **(1.)** We propose an attention based multi-task learning framework for emoji, sentiment, and emotion analysis. We leverage the

[*] First three authors have equal contributions

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

utility of sentiment and emotion and their respective intensities of the tweet to predict the emoji and vice versa. **(2.)** We crawl Twitter for additional tweets and extend the available dataset (i.e., Emoji Analysis task @ SemEval 2018). We manually add suitable emojis (😡😠🤮😨) to the extended dataset and make it rich with all the types of emojis. We further annotate the complete dataset with sentiment and emotion labels. We term the extended dataset as SEEmoji: Sentiment and Emotion aware Emoji dataset. **(3.)** We present the state-of-the-art for emoji prediction in multi-task scenario.

2 Related Work

Review of the existing research (Barbieri et al., 2018; Jin and Pedersen, 2018; Wang and Pedersen, 2018; Eisner et al., 2016; Zhou and Wang, 2017; Al-Halah et al., 2019; Felbo et al., 2017; Chen et al., 2018b; Cappallo et al., 2018; Yeh et al., 2019; Chen et al., 2018a; Cowie et al., 2001) suggests that emoji, sentiment and emotion analysis are important areas in the field of Natural Language Processing (NLP).

Emoji Analysis With the rampant usage of emoticons, the task of predicting emotions has become an important and essential task. Recently, authors in (Barbieri et al., 2017) proposed several Long Short Term Memory (LSTM) based frameworks for single label emoji prediction. In (Barbieri et al., 2018; Jin and Pedersen, 2018; Wang and Pedersen, 2018), the authors proposed a classifier for multi-lingual emoji prediction for English and Spanish languages. The authors in (Eisner et al., 2016) released emoji2vec pre-trained embeddings. As emoticons are extensively used, therefore many researchers have focused on its usage in different works such as for emoji recommendation in instant messages (Guibon et al., 2018), emoji sense disambiguation (Wijeratne et al., 2017), understanding crisis events (Santhanam et al., 2019), building emotion classifiers (Hussien et al., 2019), sentiment analysis (Al-Halah et al., 2019; Felbo et al., 2017; Chen et al., 2018b) and emotional response generation (Zhou and Wang, 2017). Lately, Ma et al. (2020) proposed transformer based network for multi-label emoji prediction.

Sentiment Analysis Sentiment analysis refers to detecting the polarity (i.e, positive, negative, or neutral) within a piece of text, be it a sentence, a paragraph, or a complete document. (Munikar et al., 2019) used BERT framework for fine-grained sentiment analysis and have shown that how effective is transformer for the NLP tasks. In other work, a document embedding using cosine similarity instead of dot product was employed for document-level sentiment analysis in (Thongtan and Phienthrakul, 2019). Sentiment classification is the task of identifying the opinion expressed in text and labeling them as positive, negative, or neutral (Medhat et al., 2014). This task has many important applications such (i) as improving the customer service by analyzing their reviews; and (ii) extracting opinions from tweets (Smailović et al., 2013), etc.

Emotion Analysis Analyzing the emotion properly also plays a significant role, like sentiment analysis, for taking better decision in many domains. Chen et al. (2018a) released a dataset taken from Friends TV series for detecting emotions in dialogues. Similarly, an attention framework was designed for identifying emotions in spoken dialog systems in (Yeh et al., 2019). Emotion classification (Cowie et al., 2001) is closely related to sentiment classification and deals with identifying the emotion in the text. However, the differences between emotion classes are much subtler than that of sentiment classes, which makes emotion classification a harder task. Recent methods have demonstrated that training a neural network jointly for both emotion and sentiment classification tasks is beneficial for both the tasks (Akhtar et al., 2018).

Our current work differentiates from the existing works on emoji prediction as we aim to leverage the sentiment and emotion and their respective intensities information for solving the problem of emoji detection in a multi-task framework and vice versa. We demonstrate through a detailed empirical evaluation that emoji detection can be improved significantly if we are successful in leveraging the knowledge of emotion and sentiment using an effective multi-task framework.

Data collection and processing: Emoji Analysis task @ SemEval 2018 (Barbieri et al., 2018) dataset consists of approx. 5.3L tweets, and each tweet is accompanied by one emoji label out of 20 emojis (🎄😜✨😎🇺🇸🔥😉😁💯😂💜😀❤️😍💙💕😘😊📷📸). These tweets were retrieved with Twitter's API and geolocalized in the United States (English). The tweets were gathered from October 2015 to February 2017. We show some of the examples in Table 2.

We hypothesize that emoji is closely related to

No.	Utterances	Emoji
1	*LoL @ West Covina, California*	😂
2	*Momma @ Disney's Magic Kingdom*	✨
3	*"A daughter is a gift of love." #family @ Vander Veer Botanical Park*	💜
4	*Free mornings spent at the beach with my girl are some of my favourite mornings #beach*	❤️
5	*Our sign is up! So awesome to see it all coming to life... Right before my eyes #makeuplounge*	✨

Table 2: Some examples from Semeval dataset.

Figure 1: Word cloud for Semeval dataset

sentiment and emotion. Sentiment analysis deals with determining the opinion (i.e., positive, negative, and neutral) expressed by a person for a topic, event, product, or a service. While, emotion analysis deals with determining the emotion displayed by a person on a topic, event, product or service (i.e., angry, disgust, fear, joy, sad, and surprise). But the emojis present in SemEval dataset are limited. The sentiment they reflect is *positive*, and the emotions displayed are *joy* and *surprise*. There are no negative sentiment emojis, for e.g. 😡 (high anger), 😠 (low anger) 🤮 (disgust), etc., are in this dataset. We show the word cloud corresponding to the SemEval dataset in Figure 1 which shows the positive nature of the dataset. To avoid this issue, we download the negative sentiment oriented tweets (approx. 2.03L) with Twitter API by using #Angry, #Disgust, #Fear, and #Sad and filter out the irrelevant tweets manually. We show the word cloud for the downloaded tweets which show their negative polarities.

Figure 2: Word cloud for extended dataset with negative emojis

We manually add one of the seven emojis 😡 (high anger), 😠 (low anger), 🤮 (disgust), 😨 (low fear), 😱 (high fear), 😢 (low sad), 😞 (high sad) as suitable for each tweet. We also show some example in Table 3.

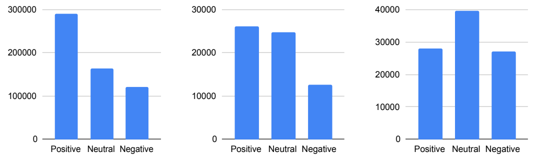
	Utterances	Emoji
1	*sooo sick of the snow ughh*	😡
2	*Damn vending machine. My skittles got stuck and i can't get them out. Can this day get any worse?*	😢
3	*People make me sick*	🤮
4	*Awww god"this fucking flu... ugh*	😠
5	*Some are just so selfish*	🤮

Table 3: Some downloaded samples with negative emojis

We then extend the SemEval dataset with these additional tweets and further annotate the complete dataset with sentiment and emotion labels (c.f. Table 1). We term the extended dataset as *SEEmoji*: Sentiment and Emotion aware Emoji dataset. We show the word cloud for the *SEEmoji* dataset which shows the positive and negative nature of the dataset.

Figure 3: Word cloud for *SEEmoji* dataset

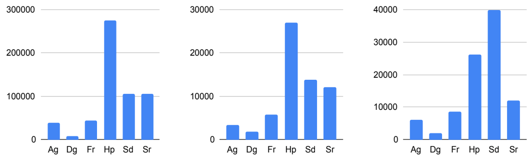
Sentiment Sentiment analysis deals with determining the polarity of the opinion expressed by a person on a topic, event, product or service. So, we consider three sentiment classes, namely *positive, negative* and *neutral* to annotate the tweet. We show some examples in Table 1. We show the overall ratio of *positive, negative* and *neutral* classes in Table 4. We also show the distribution of sentiment in terms of train set, vaild set, and test in Figure 4.



(a) *Train set.* (b) *Dev set.* (c) *Test set.*

Figure 4: SEEmoji dataset: distribution of sentiment in terms of train set, valid set, and test set.

Emotion Emotion analysis deals with determining the emotion displayed by a person on a topic, event, product or a service. We annotate each tweet with six emotion values, *viz.* angry, disgust, fear, joy, sad, and surprise. We show some example in Table 1. Table 4 shows the overall ratio of emotion labels. We also show the distribution of emotion in terms of train set, vaild set, and test in Figure 5. We divide the *SEEmoji* dataset into three sets i.e., train set, development set (dev set), and test set. We show the dataset statistics in Table 4.



(a) *Train set.* (b) *Dev set.* (c) *Test set.*

Figure 5: SEEmoji dataset: distribution of emotion in terms of train set, valid set, and test set.

Statistics	**SEEmoji Dataset**		
	Train	***Dev***	***Test***
#Tweets	575268	63589	94589
#Positive	289816	26171	28059
#Neutral	163828	24741	39496
#Negative	121624	12677	27034
#Anger	38097	3444	6174
#Disgust	7615	1756	1999
#Fear	44705	5711	8524
#Happy	270080	26024	24655
#Sad	108771	14629	41229
#Surprise	106000	12025	12008

Table 4: Dataset statistics with sentiment and emotion.

Annotation Guidelines We extend the dataset by including negative tweets in the given dataset as we have described above. We employ three graduate students highly proficient in English language with prior experience in labeling *emoji*. The guidelines for annotation, along with some examples, were explained to the annotators before starting the annotation process. Then, we annotate all the tweets with emojis. A majority voting scheme was used for selecting the final emoji label. We achieve an overall Fleiss' (Fleiss, 1971) kappa score of 0.81, which is considered to be reliable. We further annotate the sentiment and emotion labels using pre-trained models. We use TextBlob[2] for annotating sentiment and twitter-emotion-recognition[3] for annotating emotion corresponding to each tweet.

3 Proposed Methodology

In this section, we describe our proposed methodology[4]. We depict the overall architecture in Figure 6. We aim to leverage the sentiment and emotion information for solving the problem of emoji detection in a multi-task framework, and vice versa. Conneau et al. (2019) developed XLM-RoBERTa (Conneau et al., 2019), a general-purpose sentence representation and an enhanced version of mBERT and XLM ((Lample and Conneau, 2019);(Devlin et al., 2018)). XLM-RoBERTa model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

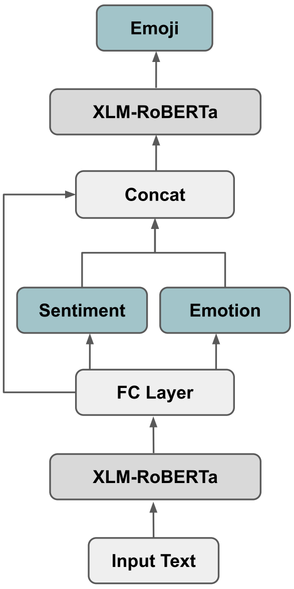


Figure 6: Proposed model

We pass the input sequence through a XLM-RoBERTa to obtain the hidden representations for emotion, sentiment. After getting the hidden representation from the linear layer, we concatenate the hidden representation and different tasks outputs. We then send this concatenated output to another XLM-RoBERTa layer to get the output for Emoji task. For finding single task result we use single XLM-RoBERTa encoder and get the output.

Multi-task loss function L1: The main objective of our loss function is to teach the model how to weight the task specific losses. For this, we adopt a principled approach to multi-task deep learning that considers the homoscedastic uncertainty[5] (Kendall

[2] https://textblob.readthedocs.io/en/dev/

[3] https://github.com/nikicc/twitter-emotion-recognition

[4] We will release the code and data.

[5] Aleatoric uncertainty that is not reliant on the input data is known as task dependant or homoscedastic uncertainty. It is not a model output, but rather a number that is constant across

et al., 2018) of each task while weighing multiple loss functions.

$$L_1 = \sum_i W_i L_i \tag{1}$$

Where i defines the different tasks (i.e. emotion, sentiment and their respective intensities). The weights are updated using back-propagation for specific losses for each tasks. For emotion and sentment we use CrossEntropyLoss and for their intensities we use MSE loss function.

For L_2 we use CrossEntropyLoss for finding best possible emoji for given tweet.

The total loss is:

$$L_{total} = L_1 + L_2 \tag{2}$$

4 Experiment results and analysis

In this section, we discuss about experimental setup, experiment results, and analysis.

4.1 Experimental Setup

We address three different tasks i.e. emoji, sentiment, and emotion analysis in a multi-task framework. We define the following experimental setups.

Emoji Classification (E^M): There are twenty seven different emojis in the SEEmoji dataset and only one emoji is associated with each tweet.

Sentiment Intensity (S_I): There are three sentiment classes associated with each tweet (i.e., negative, neutral, positive) and each sentiment value lies in the range of [-1,1].

Sentiment Classification (S_C): There are three sentiment classes associated with each tweet i.e., negative ($value < 0$), neutral ($value = 0$), and positive ($value > 0$).

Emotion Intensity (E_I): There are six emotions associated with each tweet (i.e., anger, fear, disgust, joy, sad, and surprise) and each emotion value lies in the range of [0,1].

Emotion Classification (E_C): We, at first, find the maximum value among six emotions then put one at the maximum place and zero for rest places.

We implement our proposed model in PyTorch[6], a Python-based deep learning library. We perform *grid search* to find the optimal hyper-parameters (c.f. Table 5). As evaluation metrics, we use accuracy and F1-score for the classification problems, while for the intensity prediction task, we compute the mean square error *(MSE)*, mean absolute error *(MAE)*, pearson correlation scores (P-corr), and cosine similarity *(Cos)* to show the performance of our proposed model. We use *Adam* as an optimizer.

Parameters	SEEmoji Dataset
XLM-RoBERTa	'xlm-roberta-base' ,Dropout=0.05
FC	2*768, Dropout=0.05
Activations	*ReLu* as activation for our model
Output	Softmax (E^M, S_C, E_C), tanh (S_I), & sigmoid (E_I)
Optimizer	Adam (lr=0.001)
Model Loss	Cross-entropy (Classification) & MSE (Intensity)
Batch	32
Epochs	50

Table 5: Hyper-parameters for our experiments where *N*, *D*, S_C, S_I, E_C, and E_I stands for #neurons, dropout, sentiment classification, sentiment intensity, emotion classification, and emotion intensity, respectively.

We use *Softmax* as a classifier for emoji, sentiment and emotion classification, and optimize the *cross entropy* loss. For sentiment and emotion intensity, we use *tanh* and *sigmoid* activation, respectively, on the output layers, and optimize the mean-squared-error *(MSE)* loss.

Results and Analysis. We evaluate our proposed approach for all the possible combinations of the tasks which are as follows:

Uni task learning (UTL): A separate model is trained for all different dimensions i.e., emoji classification (Emoji), sentiment classification (S_C), sentiment intensity (S_I), emotion classification (E_C), and emotion intensity (E_I). **Dual task learning (DTL):** Two tasks (i.e., emoji and sentiment or emoji and emotion etc.) are trained together (c.f. DTL in Table 6). **Tri task learning (TTL):** Three tasks (i.e., emoji, sentiment, and emotion etc) are trained together (c.f. TTL in Table 6).

Emoji Classification (E^M) We show the emoji classification results in Table 6. For *TTL*, our model achieves 7.99% and 4.77% improvement in F1-score compared to *UTL* and *DTL*, respectively. We see similar improvement in accuracy also. We observe that the proposed approach yields better performance for the *TTL* than the *DTL* and *UTL*. This improvement implies that our proposed hypothesis is correct and very effective. We also present the bar-chart to show the improvement in Figure 7.

Sentiment Classification (S_C) We show the sentiment classification results in Table 7. For *TTL*, our model achieves 4.68% and 2.93% improvement in F1-score compared to *UTL* and *DTL*, respectively. We see similar improvement in accuracy also. We

all input data and changes between tasks. As a result, it is known as task-dependent uncertainty.

[6] https://pytorch.org/

Tasks		F1-score	Accuracy
UTL	E^M	45.30	48.23
DTL	S_C + E^M	47.26	48.63
	S_I + E^M	48.52	50.28
	E_C + E^M	46.45	49.29
	E_I + E^M	46.12	51.32
TTL	S_C + E_C + E^M	**53.29**	**55.86**
	S_C + E_I + E^M	50.39	51.32
	S_I + E_C + E^M	50.37	52.36

Table 6: Emoji classification results

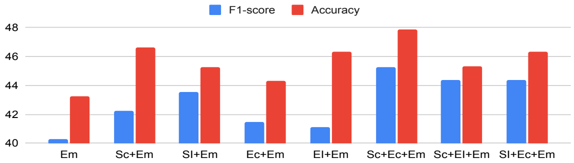


Figure 7: Bar chart for emoji classification which shows the improvement over UTL and DTL.

observe that the proposed approach yields better performance for the *TTL* than the *DTL* and *UTL*. Thus, we can say emoji and emotion class E_C are helping to sentiment class (S_C). We also present the bar-chart to show the improvement in Figure 9a.

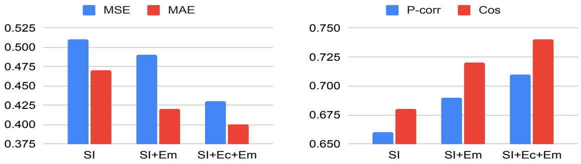
Tasks		F1-score	Accuracy
UTL	S_C	91.93	93.95
DTL	S_C + E^M	94.61	95.54
TTL	S_C + E_C + E^M	**96.61**	**97.54**

Table 7: Sentiment classification results

Sentiment Intensity (S_I) We show the sentiment intensity results in Table 8. We report the results for metrics[7] MSE, MAE, P-corr, and cos. We observe that the proposed approach yields better performance for the *TTL* than the *DTL* and *UTL*. We present the bar-chart to show the improvement in Figure 8.

Tasks		MSE	MAE	P-corr	Cos
UTL	S_I	0.51	0.47	0.66	0.68
DTL	S_I + E^M	0.49	0.42	0.69	0.72
TTL	S_I + E_C + E^M	**0.43**	**0.40**	**0.71**	**0.74**

Table 8: Sentiment intensity results



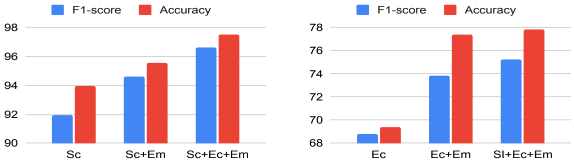
(a) *MSE and MAE.* (b) *P-corr and cos.*

Figure 8: Bar chart for sentiment intensity which shows the improvement over UTL and DTL.

Emotion Classification (E_C) We show the emotion classification results in Table 9. Similar to sentiment classification, we observe that the proposed approach yields better performance for the *TTL* than the *DTL* and *UTL*. We present the bar-chart to show the improvement in Figure 9b.

Tasks		F1-score	Accuracy
UTL	E_C	68.80	69.37
DTL	E_C + E^M	73.82	77.39
TTL	S_C + E_C + E^M	**75.23**	**77.84**

Table 9: Emotion classification results.



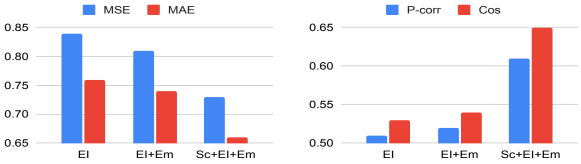
(a) *Sentiment Classification.* (b) *Emotion Classification.*

Figure 9: Bar chart for sentiment intensity which shows the improvement over UTL and DTL.

Emotion Intensity (E_I) We show the emotion intensity results in Table 10. Similar to sentiment intensity, we observe that the proposed approach yields better performance for the *TTL* than the *DTL* and *UTL*. We present the bar-chart to show the improvement in Figure 10.

Tasks		MSE	MAE	P-corr	Cos
UTL	E_I	0.84	0.76	0.51	0.53
DTL	E_I + E^M	0.81	0.74	0.52	0.54
TTL	S_C + E_I + E^M	**0.73**	**0.66**	**0.61**	**0.65**

Table 10: Emotion intensity results



(a) *MSE and MAE.* (b) *P-corr and cos.*

Figure 10: Bar Chart for Emotion Intensity which shows the improvement over UTL and DTL.

[7] Please note that while higher values of Pearson score and Cosine similarity are the indicators of better performance, lower values of mean-squared-error (MSE) and mean-absolute-error (MAE) correspond to the better performance

	Tweets		Emoji	Sentiment		Emotion						
							Intensity					
			Class	Class	Intensity	Class	Ag	Dg	Fr	Jy	Sd	Sr
T_1	*Hates prank calls. Especially when they're from pple who sound like terrorists.*	Actual	🤮	**Neg**	**-0.23**	**Dg**	**0.21**	**0.53**	**0.00005**	**0.001**	**0.114**	**0.143**
		UTL (S, E, E^M)	😠	*Neg*	*-0.051*	Sr	*0.012*	*0.316*	*0.0*	*0.00002*	*0.10*	*0.13*
		DTL (S_C, E^M)	😠	Neg	-	-	-					
		DTL (S_I, E^M)	😡	-	*-0.010*	-	-					
		DTL (E_C, E^M)	😢	-	-	Ag	-					
		DTL (E_I, E^M)	😞	-	-	-	*0.11*	*0.151*	*0.0000001*	*0.00022*	*0.06*	*0.112*
		TTL (S_C, E_C, E^M)	🤮	Neg	-	Dg	-					
		TTL (S_C, E_I, E^M)	🤮	Neg	-	Dg	*0.18*	*0.41*	*0.0*	*0.000031*	*0.10*	*0.1305*
		TTL (S_I, E_C, E^M)	🤮	Neg	-0.16	Dg	-					
T_2	*One of this weekends weddings. Love red flowers on white cakes! #katscakesnola #nolawedding*	Actual	😍	**Pos**	**0.1666**	**Jy**	**0.00007**	**0.00002**	**0.0005**	**0.88**	**0.001**	**0.11**
		UTL (S, E, E^M)	😁	*Pos*	*0.0042*	Sr	*0.00002*	*0.016*	*0.0*	*0.42*	*0.000023*	*0.024*
		DTL (S_C, E^M)	😀	Pos	-	-	-					
		DTL (S_I, E^M)	😀	-	*.049*	-	-					
		DTL (E_C, E^M)	😘	-	-	Sr	-					
		DTL (E_I, E^M)	😀	-	-	-	*0.0021*	*0.0013*	*0.0003*	*0.49*	*0.06*	*0.0058*
		TTL (S_C, E_C, E^M)	😍	Pos	-	Jy	-					
		TTL (S_C, E_I, E^M)	😍	Pos	-	Jy	*0.0*	*0.000011*	*0.00001*	*0.62*	*0.00089*	*0.105*
		TTL (S_I, E_C, E^M)	😍	Pos	0.067	Jy	-					
T_3	*Coach made me shave. That made me mad. haa*	Actual	😡	**Neg**	**-0.2125**	**Ag**	**0.962**	**0.003**	**.01**	**0.003**	**0.01**	**0.009**
		UTL (S, E, E^M)	😂	*Pos*	*0.025*	Sr	*0.42*	*0.001*	*0.003*	*0.0002*	*0.20*	*0.0051*
		DTL (S_C, E^M)	😀	Pos	-	-	-					
		DTL (S_I, E^M)	😊	-	*-0.031*	-	-					
		DTL (E_C, E^M)	😁	-	-	Sr	-					
		DTL (E_I, E^M)	😀	-	-	-	*0.31*	*0.021*	*0.003*	*0.0012*	*0.0013*	*0.00058*
		TTL (S_C, E_C, E^M)	😡	Neg	-	Ag	-					
		TTL (S_C, E_I, E^M)	😡	Neg	-	Ag	*0.829*	*0.00121*	*0.05*	*0.002*	*0.0074*	*0.0038*
		TTL (S_I, E_C, E^M)	😡	Neg	-0.167	Ag	-					
T_4	*I can't wait to see this cutie in a couple of days. I miss him so much. #mybaeisinthebay @user*	Actual	❤️	**Pos**	**0.431**	**Sd**	**0.003**	**0.0001**	0.007	0.09	0.851	0.12
		UTL (S, E, E^M)	😊	*Pos*	*0.31*	Jy	*0.001*	*0.00051*	*.23*	*0.05*	*0.34*	*0.092*
		DTL (S_C, E^M)	😘	Pos	-	-	-					
		DTL (S_I, E^M)	😘	-	*0.351*	-	-					
		DTL (E_C, E^M)	😢	-	-	Sd	-					
		DTL (E_I, E^M)	😞	-	-	-	*0.0015*	*0.0019*	*0.006*	*0.022*	*0.46*	*0.02*
		TTL (S_C, E_C, E^M)	❤️	Pos	-	Sd	-					
		TTL (S_C, E_I, E^M)	😞	Pos	-	Dg	*0.0021*	*0.0041*	*0.03*	*0.0045*	*0.30*	*0.11*
		TTL (S_I, E_C, E^M)	😍	Pos	0.067	Jy	-					
T_5	*@user I wanna own your business, but I'm 16 and have no money.*	Actual	✨	**Pos**	**0.376**	**Sd**	**0.0004**	**0.002**	**0.00005**	**0.001**	**0.842**	**0.142**
		UTL (S, E, E^M)	😉	Pos	*0.0039*	Jy	*0.0002*	*0.00306*	*.00001*	*0.001*	*0.42*	*0.11*
		DTL (S_C, E^M)	😀	Pos	-	-	-					
		DTL (S_I, E^M)	😊	-	*0.142*	-	-					
		DTL (E_C, E^M)	😠	-	-	Dg	-					
		DTL (E_I, E^M)	😢	-	-	-	*0.00021*	*0.00047*	*0.0*	*0.0002*	*0.392*	*0.128*
		TTL (S_C, E_C, E^M)	😠	Pos	-	Sd	-					
		TTL (S_C, E_I, E^M)	😢	Pos	-	Sd	*0.00037*	*0.0011*	*0.00002*	*0.0013*	*0.51*	*0.09*
		TTL (S_I, E_C, E^M)	😞	Pos	0.29	Sd	-					

Table 11: **Qualitative analysis of the Uni task learning (UTL), Dual task learning (DTL) and Tri task learning (TTL) frameworks.** Few error cases where Tri task learning framework performs better than the uni-task and dual task framework. We also some examples where Tri task learning does not work well with reason. **Ag:** Anger, **Dg:** Disgust, **Fr:** Fear, **Jy:** Joy, **Sd:** Sad and **Sr:** Surprise. The *red colored text* shows error in classification, while the *blue colored text* reflects predicted intensity values.

5 Error Analysis

In this section, we present the error analysis of our proposed multitask framework. We stated earlier that emoji, sentiment, and emotion are highly related to each other. To show the effect of these tasks on each other, we take some examples from SEEmoji dataset (c.f. Table 11). First tweet (T_1) in Table 11 "*Hates prank calls. Especially when they're from pple who sound like terrorists*" has emoji 🤮 with negative sentiment and disgust emotion. Our *TTL* predicts the emoji correctly while *DTL* fails to predict the correct emoji and emotion. We observe that sentiment and emotion together help to predict the correct emoji. In other words, we can say sentiment and emotion also help each other. While in some tweets, TTL fails to predict correct emoji, e.g., fifth tweet in Table 11, "@user I wanna own your business, but I'm 16 and have no money." has emoji ✨ but *TTL* fails to predict 😞 emoji because of emotion. *TTL* predicts the correct emotion as sad and *w.r.t.* sad *TTL* predicts the sad emoji as well. There are twenty seven emojis and only six emotions which is 4.5 emojis/emotion. This is the reason behind when *TTL* does not predict the correct emoji but predicts sentiment and emotion correctly.

6 Conclusion

In this paper, we have proposed an effective deep learning-based multi-task model to simultaneously solve all the three problems, *viz.* emoji analysis, sentiment analysis, and emotion analysis. We used the already available dataset (Emoji Analysis task @ SemEval 2018) which contains, along with tweets, emojis that convey positive sentiment. To make the dataset rich with all types of emojis, we extended it with additional tweets and, accordingly, manually add emojis, as suitable for each tweet. We further annotated the complete dataset with sentiment and emotion labels. We term the extended dataset as SEEmoji: Sentiment and Emotion aware Emoji dataset. Empirical results on SEEmoji dataset indicates that the proposed multi-task framework yields better performance over the single-task learning. During our analysis, we found that more than one emoji is possible for a given tweet. So, we will try to make a group of emojis (multi-emoji) corresponding to each tweet and perform multi-label emoji prediction with sentiment and emotion.

References

Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2018. A multi-task ensemble framework for emotion, sentiment and intensity prediction. *arXiv preprint arXiv:1808.01216*.

Ziad Al-Halah, Andrew Aitken, Wenzhe Shi, and Jose Caballero. 2019. Smile, be happy:) emoji embedding for visual sentiment analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? *arXiv preprint arXiv:1702.07285*.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.

Spencer Cappallo, Stacey Svetlichnaya, Pierre Garrigues, Thomas Mensink, and Cees GM Snoek. 2018. New modality: Emoji challenges in prediction, anticipation, and retrieval. *IEEE Transactions on Multimedia*, 21(2):402–415.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018a. Emotionlines: An emotion corpus of multi-party conversations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018b. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 117–125.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gaël Guibon, Magalie Ochs, and Patrice Bellot. 2018. Emoji recommendation in private instant messages. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1821–1823.

Wegdan Hussien, Mahmoud Al-Ayyoub, Yahya Tashtoush, and Mohammed Al-Kabi. 2019. On the use of emojis to train emotion classifiers. *arXiv preprint arXiv:1902.08906*.

Shuning Jin and Ted Pedersen. 2018. Duluth urop at semeval-2018 task 2: Multilingual emoji prediction with ensemble learning and oversampling. *arXiv preprint arXiv:1805.10267*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Emoji prediction: Extensions and benchmarking. *arXiv preprint arXiv:2007.07389*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE.

Sashank Santhanam, Vidhushini Srinivasan, Shaina Glass, and Samira Shaikh. 2019. I stand with you: Using emojis to study solidarity in crisis events. *arXiv preprint arXiv:1907.08326*.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2013. Predictive sentiment analysis of tweets: A stock market application. In *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88. Springer.

Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414.

Zhenduo Wang and Ted Pedersen. 2018. Umdsub at semeval-2018 task 2: Multilingual emoji prediction multi-channel convolutional neural network on subword embedding. *arXiv preprint arXiv:1805.10274*.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. Emojinet: An open service and api for emoji sense discovery. In *Eleventh International AAAI Conference on Web and Social Media*.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6685–6689. IEEE.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*.

Annotation and Multi-modal Methods for Quality Assessment of Multi-party Discussion

Tsukasa Shiota
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka JAPAN

Kazutaka Shimada
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka JAPAN
shimada@ai.kyutech.ac.jp

Abstract

Discussion quality assessment tasks have recently attracted significant attention in natural language processing. However, there have been few studies on challenging such tasks, with a focus on synchronous discussions. In this study, we annotate quality scores to each discussion in an existing multi-modal multi-party discussion corpus. Furthermore, we propose some quality assessment methods with multi-modal inputs. As the results show, attention-based long short-term memory (LSTM) with multi-modal inputs produces the best performance for the "Effectiveness" criterion whereas text information has an important role in the "Reasonableness."

1 Introduction

In recent years, problem-based and cooperative learning have been attracting attention as a means of skills training, such as communication skills, in education. One educational training approach is a group discussion, which involves debate and consensus-building. Introducing this learning approach to a classroom requires a great deal of effort to evaluate and provide feedback on the abilities and achievements of all groups and individuals from various perspectives because several discussion groups usually exist in a single class at the same time. Furthermore, an assessment is a difficult task because there are no correct answers regarding discussions in general. Moreover, quantitative and objective evaluations are difficult. Therefore, an automatic assessment, such as a visualization of the discussion state and a judgment of the discussion score, is a desirable and valuable task for education, examinations through discussion, and so forth. It will be possible to reduce the burden of evaluation activities on the evaluators.

One of the educational applications in natural language processing is automated essay scoring (AES) (Ke and Ng, 2019) as an argument quality evaluation. However, the structure of a spoken discussion, which is our target in this paper, is not as clear as that of a written discussion. In addition, in spoken discussions, both verbal and non-verbal information have important roles in understanding and evaluating the discussion. Mukawa et al. (2018) have reported that non-verbal features, such as gestures and an interval of utterances, have a powerful effect on group discussions.

In this study, we annotate several quality assessment criteria and scores to a multi-modal multi-party discussion corpus. The language used is Japanese, and the corpus is freely available[1]. In addition, we propose the use of machine-learning-based methods, such as a support vector machine (SVM) and neural networks, and then evaluate the methods using multi-modal inputs. In the experiment, we discuss the relationships between the assessment criteria and input modalities.

2 Related work

There are some dialogue and meeting corpora (Carletta, 2007; Janin et al., 2003). Some face-to-face discussion corpora have been also developed. Zhang et al. (2016) have constructed a corpus through a competitive debate format. They reported that their method predicts the winner of each debate at a rate of approximately 60%. Hayashi et al. (2015) have developed a group discussion interaction corpus to evaluate five communication skills. This corpus contains not only transcriptions but also speech, gaze, head motions, and poses using certain devices. Olshefski et al. (2020) have constructed a discussion tracker corpus in an educational environment. The corpus consists of 29 multi-party discussions. Yamamura et al. (2016) have constructed a corpus for a discussion summarization. The corpus consists of 9 discussions by four participants.

[1] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html#kyutechDB

As mentioned in Section 1, many studies and corpora of asynchronous and written texts exist, such as essay writing (Ke and Ng, 2019). Some researchers have recently studied interactions between participants during discussions. For example, Okada et al. (2016) have annotated communication skill scores on the MATRICS corpus (Nihei et al., 2014). They also proposed a multi-modal prediction model for such a task. In addition, Avci and Aran (2016) and Murray and Oertel (2018) have proposed performance prediction models by using features extracted from the states of the discussions and the participants. In this paper, we also introduce multi-modal features to our quality assessment method.

3 Dataset

Our purpose in this paper is to assess a quality of a multi-party discussion. For the purpose, we need a discussion corpus. In this paper, we utilize the corpus that was constructed by (Shiota and Shimada, 2020), namely the Kyutech Debate corpus. It is freely available on their website[2]. This section describes their corpus first and then our annotation process for our purpose.

In the Kyutech Debate corpus, two people in a group first debated an issue from both positive and negative standpoints, and the two groups then came to a consensus through compromise. The first (debate) and second (consensus-building) parts were each 20 min in length. The discussions of five groups were recorded, with 200 min of discussions as a whole. The corpus consisted of 7,449 utterances that were transcribed[3], body key-points determined by OpenPose[4], facial landmarks determined by OpenFace[5], and the speech features analyzed using Surfboard[6].

In this paper, we newly add quality assessment scores for the corpus. In general, participants need to discuss various topics in the case of debating/consensus-building of an issue. Hence, we extract topic-based segments (hereinafter referred to as "discussion segments") and regard them as the target units of a quality assessment. We referred to the topic segmentation manual (Xu et al., 2005) of the AMI corpus, which is a popular conversation corpus. As a result, we obtained 178 segments from the Kyutech Debate corpus.

Next, we created criteria based on the theory of computational quality assessment of natural language arguments (Wachsmuth et al., 2017b) and conducted a grading process. According to the classification defined by the above study, the quality of an argument can be evaluated through two main criteria, "Reasonableness (Re)" and "Effectiveness (Ef)," and their sub-criteria. The sub-criteria of Re are "Global Acceptability (GA)," "Global Relevance (GR)," and "Global Sufficiency (GS)." The sub-criteria of Ef are "Credibility (Cr)," "Emotional appeal (Em)," "Clarity (Cl)," "Appropriateness (Ap)," and "Arrangement (Ar)." Table 1 provides a description of each criterion based on the previous study.

Three workers, who were graduate students and not related to this work in our laboratory, were assigned to each discussion segment. Given the transcription and video data of a discussion segment, they judged the quality of each segment on the basis of the more detailed explanation provided in Table 1. The first step was to rate each sub-criterion as low (L), middle (M), or high (H), and then determine the score of the main criteria on the basis of the score distribution of the sub-criteria: very low (VL), L, M, H, or very high (VH).

The reliability of the annotated main and sub-criteria was confirmed by calculating the agreement rate. Table 2 shows Krippendorff's α coefficient for each criterion. This coefficient is a continuous value of between -1 and 1, which can be used to calculate the rate of agreement for all scales. The values in the table are not always high. The result denotes that the annotation task is inherently difficult. As a similar study, Wachsmuth et al. (2017a) also reported an annotation process and the result using the same scheme for the written text. In their study, the Klippendorff α of crowd workers ranged from -0.27 to 0.53. In other words, the result was also low. Moreover, the values of some criteria by Wachsmuth et al. (2017a) dropped below zero. On the other hand, such results did not appear on our annotation. Therefore, our annotated data contain a better point than the previous study. In other words, our data are a better-than-random chance although the previous work contained a result that was less than a ran-

[2] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

[3] Transcription units were based on a 0.2 seconds interval.

[4] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[5] https://github.com/TadasBaltrusaitis/OpenFace

[6] https://github.com/novoic/surfboard

Criterion	Explanation
Re	Does the argumentation satisfy GA, GR, and GS?
GA	Does the target audience accept both the consideration of the stated arguments regarding the issue and the way in which they are stated?
GR	Does it contribute to the resolution of the issue?
GS	Does it adequately rebut the counterarguments by properly anticipating them?
Ef	Does the argumentation satisfy Cr to Ar?
Cr	Does it convey arguments and is it similar in such a way that it makes the author worth considering?
Em	Were emotions elicited to make the target audience more open to the author's arguments?
Cl	Was the argument correct and widely unambiguous?
Ap	Did the language used support the credibility?
Ar	Were the issue, arguments, and their conclusion presented in the correct order?

Table 1: Definitions of the Quality Dimensions, based on Wachsmuth's study.

Re	GA	GR	GS	-	-
0.151	0.087	0.029	0.128	-	-
Ef	Cr	Em	Cl	Ap	Ar
0.135	0.032	0.038	0.017	0.076	0.155

Table 2: Krippendorff's α of each criterion.

dom chance. Although this annotation is a complicated task, it is necessary to improve the agreement as one future work. As one of our contributions, we will open the annotated data on the web.

4 Quality Assessment Method

In this section, we describe our quality assessment method for the dataset introduced in Section 3. First, we define the quality assessment task and then propose four models based on the SVM and neural networks.

4.1 Task Definition

A discussion segment S consists of a sequence of utterance vectors $U = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N\}$. Here, N is the number of utterances in a segment, and $\boldsymbol{u}_i$ is a vector of the i-th utterance in S. The task in this paper is to predict the class labels of each criterion from a sequence U with verbal and non-verbal information. Here, the class labels are L, M, and H, as described in Section 3. Owing to the limited number of instances that belong to each class, VL is merged with L, and VH is merged with H. In other words, the task is a classification task with three class labels for the criteria in Section 3.

Here, $\boldsymbol{u}_i$ is expressed as follows:

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{t}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \qquad (1)$$

where $[\cdot;\cdot]$ denotes the concatenation of the vectors.

In addition, $\boldsymbol{sp}_i$ denotes whether the speaker of the i-th utterance is different from the speaker of the $i-1$-th utterance. In other words, it is a binary feature, i.e, the speaker is the same (0) or different (1).

Moreover, $\boldsymbol{t}_i$ is a vector from text information of the i-th utterance. For the $\boldsymbol{t}_i$, we use BERT (Devlin et al., 2019). We apply the CLS token (768 dimensions) on the 11th layer from a Japanese BERT developed by Tohoku University[7].

Here, $\boldsymbol{b}_i$ is a vector from the body information of the i-th utterance. It consists of the average and standard deviation of (x, y) values of the nose, neck, right shoulder, right elbow, right wrist, right eye, right ear, left shoulder, left elbow, left wrist, left eye, and left ear from OpenPose (a total of 48 dimensions).

In addition, $\boldsymbol{f}_i$ is a vector from facial information of the i-th utterance. It consists of the average and standard deviation of the facial and eye points (x, y), gaze direction, head location, and head direction. In addition, it contains the presence of facial action units (AUs). OpenFace extracts these values, and the number of dimensions is 586.

Finally, $\boldsymbol{a}_i$ is a vector from audio information of the i-th utterance. It consists of the minimum, maximum, average, and standard deviation of 13 MFCC, the RMS, the fundamental frequency, and the spectral centroid. In addition, it contains the Jitter and Shimmer values. Surfboard extracts these values, and the number of dimensions is 72.

[7] https://github.com/cl-tohoku/bert-japanese

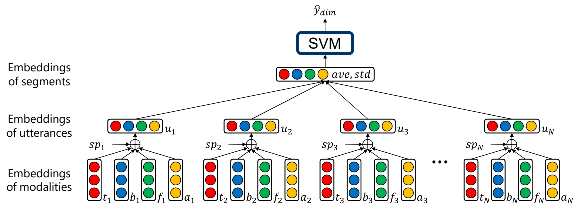


Figure 1: Method based on SVM.

4.2 SVM

As one of the simplest models, we apply an SVM (Vapnik, 2013) to the task. Because an SVM cannot handle sequence information directly, we compute the average and standard deviation of each vector in the time sequence and use the values as the vector of each discussion segment. We estimate each quality assessment label $\hat{y}_{dim}$ by using the model with the vector. Figure 1 shows an overview of this method.

4.3 LSTM

As mentioned above, the SVM-based method cannot handle the utterance sequence information well. Therefore, as a suitable model for sequence information, we use LSTM for this task.

Given an input $\boldsymbol{u}_i$, the units of LSTM are computed as follows:

$$\boldsymbol{h}_i = LSTM(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1}) \tag{2}$$

After the computation for all utterances, LSTM obtains the final state of a discussion segment $\boldsymbol{h}_N$. In this paper, we regard $\boldsymbol{h}_N$ as the embeddings of the discussion segment. We calculate a probability distribution $\hat{Y}_{dim}$ using the softmax function.

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N + \boldsymbol{b}_s), \tag{3}$$

where $\boldsymbol{W}_s$ and $\boldsymbol{b}_s$ are parameters in the learning process, respectively. In addition, $\mathrm{softmax}()$ is the softmax function. Finally, we select the label with the maximum probability ($\hat{y}_{dim}$).

$$\hat{y}_{dim} = \underset{y_{dim}}{\arg\max}\, \hat{Y}_{dim} \tag{4}$$

4.4 Attention-based LSTM

By using the LSTM, we can capture the sequence information of the utterances. However, discussion segments often contain non-important utterances for a quality assessment task, e.g., a nod. Therefore, we introduce attention mechanisms to the LSTM-based method, such as the models from (Wang et al., 2016) and (Zhou et al., 2016).

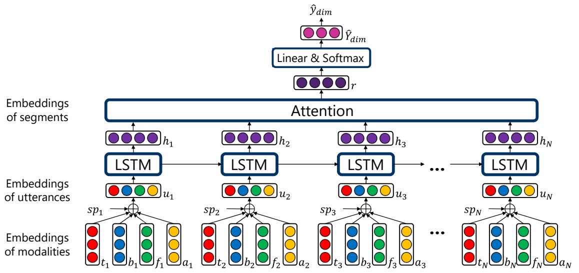


Figure 2: Attention-based LSTM.

First, in the same way as the LSTM-based approach, we compute $\boldsymbol{h_i}$ of i. We then compute the weight a_i of $\boldsymbol{h_i}$ by

$$m_i = \boldsymbol{\omega}^T \tanh(\boldsymbol{h}_i), \tag{5}$$

$$a_i = \frac{\exp(m_i)}{\sum\limits_{j=1}^{N} \exp(m_j)}, \tag{6}$$

where $\boldsymbol{\omega}^T$ is a parameter, and $\exp()$ is the exponent function. Next, we obtain the final state $\boldsymbol{h}^*$ by using the summation of hidden layers $\boldsymbol{h}_i$ weighted by a_i.

$$\boldsymbol{r} = \sum_{i=1}^{N} a_i \boldsymbol{h}_i, \tag{7}$$

$$\boldsymbol{h}^* = \tanh(\boldsymbol{r}), \tag{8}$$

where $\tanh()$ is the hyperbolic tangent function. We regard $\boldsymbol{h}^*$ as the embeddings of the segment and calculate $\hat{Y}_{dim}$.

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}^* + \boldsymbol{b}_s) \tag{9}$$

Finally, we select the label with the maximum probability ($\hat{y}_{dim}$). Figure 2 shows an overview of this method.

4.5 Hierarchical LSTM

By using an LSTM and attention mechanisms, we can handle the state of an utterance sequence. However, $\boldsymbol{t}_i$ does not directly handle the word sequence in an utterance. We therefore incorporate word sequence information with the LTSM-based model similarly to the approach by (Tran et al., 2017).

We compute $\boldsymbol{h}_{i,j}^{Uttr}$ with $\boldsymbol{w}_i$ as follows:

$$\boldsymbol{h}_{i,j}^{Uttr} = LSTM^{Uttr}(\boldsymbol{w}_{i,j}, \boldsymbol{h}_{i,j-1}^{Uttr}, \boldsymbol{c}_{i,j-1}^{Uttr}), \tag{10}$$

$$\boldsymbol{w}_i = \boldsymbol{h}_{i,M_i}^{Uttr}, \tag{11}$$

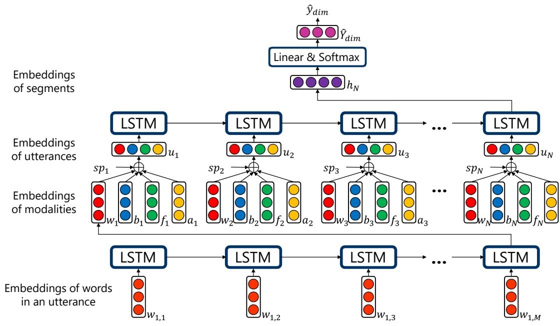


Figure 3: Hierarchical LSTM.

where $LSTM^{Uttr}()$ is an encoder of the sequence of word vectors, and $\boldsymbol{w}_{i,j}$ is the vector of the j-th word in the i-th utterance in a segment S. The vector is also extracted from the 11th layer of BERT. Here, $\boldsymbol{h}_{i,j}^{Uttr}$ is the hidden layer of the LSTM at (i, j), $\boldsymbol{c}_{i,j}^{Uttr}$ is the memory cell of the LSTM at (i, j), and M_i is the number of words in the i-th utterance. Hence, the vector $\boldsymbol{u}_i$ of this method is as follows:

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{w}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \qquad (12)$$

After that, similarly to the LSTM-based method described in Section 4.3, we also obtain the final state $\boldsymbol{h}_i^{Hier}$, $\hat{Y}_{dim}$, and the class label with the maximum probability.

$$\boldsymbol{h}_i^{Hier} = LSTM^{Hier}(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}^{Hier}, \boldsymbol{c}_{i-1}^{Hier}) \qquad (13)$$

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N^{Hier} + \boldsymbol{b}_s) \qquad (14)$$

Figure 3 shows an overview of this method.

5 Experiment

5.1 Setting

As mentioned in Section 4.1, we merged the VH class with H and the VL class with L. Hence, the task in this paper is a three-class classification, namely, L, M, and H. The targets of the classification are the two criteria with relatively high Krippendorff α values listed in Table 2: Re (Reasonableness) and Ef (Effectiveness). The statistics are shown in Table 3.

We applied the L2 norm cross-entropy loss as the loss function for the neural network-based methods. We used the SGD (Bottou, 1991) with Momentum (Qian, 1999) ($\alpha = 0.95$) as the optimizer. For the hyperparameters, the size of hidden layers was 500 dimensions, the batch size was 32, the number of epochs was 50, the learning rate was 0.01, the drop-out rate was 0.2, and the decay factor was 0.001.

Criterion	L	M	H
Re	13	89	76
Ef	9	97	72

Table 3: Distribution of each class of two target criteria.

Our dataset is small, namely, 178 segments from 10 discussions. We divided the dataset into eight discussions for the training, one discussion for the development, and one discussion for the test. We then evaluated each method based on a 10-fold cross-validation of the discussion level. We calculate the average F-scores for each criterion, i.e., Re and Ef, based on the cross-validation. For the robustness of the results, we conducted this evaluation five times and then calculated the average values of the five evaluations.

5.2 Results and Discussion

Table 4 shows the experiment results. Here, T, B, F, and A denote the text, body information, facial information, and audio information modalities. The combination of each letter denotes the combination of modalities. For example, TB denotes the combination of text and body information as the input of each method. Hence, TBFA, on the left-most side of the table, denotes the method with all modalities. Boldface denotes the best score among the modality combinations. The underlined values denote the best values of uni-modal, bi-modal, and multi-modal inputs. For example, 0.398, 0.459, and 0.399 are the best scores of TB, TF, and TA, namely, bi-modal inputs. The best score of the bi-modal setting is 0.459 by the TF input. The scores with $*$ denote the best scores for each criterion.

For the Re criterion, multi-modal inputs were not always effective for the classification. However, there were no significant changes in the results when the input modalities were expanded. The best score was 0.459, achieved by hierarchical LSTM (H-LSTM) with text and facial information. However, the difference between H-LSTM and SVM with text only was slight (0.008). Moreover, H-LSTM is a method that can handle word information directly, as compared with LSTM and attention-based LSTM (A-LSTM). Here, recall that the Re criterion consists of the acceptability,

Criteria	Model	F1-Score							
		T	TB	TF	TA	TBF	TBA	TFA	TBFA
Re	SVM	**0.451**	0.338	0.337	0.343	0.333	0.317	0.320	0.340
	LSTM	0.387	**0.398**	0.392	0.380	0.410	0.379	0.388	0.360
	A-LSTM	0.412	0.392	0.387	**0.399**	0.371	0.359	**0.398**	0.398
	H-LSTM	0.359	0.354	**0.459****	0.388	**0.415**	**0.391**	0.370	**0.405**
Ef	SVM	**0.459**	0.382	0.383	0.436	0.384	0.392	0.406	0.379
	LSTM	0.428	**0.478**	**0.438**	**0.476**	**0.467**	**0.472**	**0.486**	0.435
	A-LSTM	0.433	0.470	0.426	0.468	0.450	0.396	0.444	**0.490***
	H-LSTM	**0.459**	0.433	0.416	0.379	0.414	0.440	0.431	0.451

Table 4: Experiment results of four methods with a combination of four modalities.

relevance, and sufficiency of the discussions. In other words, it is related to the content of each discussion. Therefore, non-verbal information is less likely to contribute to improving the accuracy. From these results, we concluded that text information is the most important factor for the Re criterion.

For the Ef criterion, the combination of modalities improved the F-scores except for the SVM-based method. The best F-score was produced by A-LSTM with all modalities (0.490). The Ef criterion is based on the receivers' emotions during the discussions, such as credibility and emotional appeal. In addition, it contains clarity and appropriateness in the discussion. In general, we utilize eye contact (addressing) and body language to clearly convey a message and elicit sympathy. In other words, not only text but also actions, expressions, and the tone of voice of the speakers have an important role for the Ef criterion. From these results, we conclude that incorporating both verbal and non-verbal modalities leads to an improvement of the estimation of this criterion.

One simple method for predicting the label of a criterion is to use the majority label. In this dataset, this is label M for both criteria (Re and Ef) from Table 3. However, note that the distribution of labels in each discussion is uniform. In other words, there is a situation in which most of the labels in a discussion are H, although another discussion contains as many instances of label M as label H. In fact, the F1-scores of the majority selection based on the same calculation approach described in Section 5.1 were 0.333 for Re and 0.384 for Ef[8]. These values were lower than most of the F-scores in Table 4. This result shows the effectiveness of our methods.

6 Conclusions

In this paper, we annotated quality assessment scores for an automatic discussion evaluation to a multi-party conversation corpus. We proposed four machine learning methods for the task: SVM-based, LSTM-based, attention-based LSTM, and hierarchical LSTM methods. We used not only text but also non-verbal information, namely, multi-modal inputs.

We evaluated the methods using a 10-fold cross-validation for two criteria at the discussion level, namely, Re and Ef, in the corpus. For Re, the hierarchical LSTM with text and facial information obtained the best F-score. In addition, the SVM with only text information obtained a good result. For this criterion, text information has the most important role because it is related to the content of the discussions. For Ef, the attention-based LSTM with all modalities produced the best F-score. For this criterion, various inputs are essentially suitable because it is related to the impression and emotion of the speakers and receivers in the discussion. However, the F1-scores are insufficient (0.459 for Re and 0.490 for Ef). Improving the method using other information, such as knowledge graphs (Al-Khatib et al., 2020), is an important area of future work.

We annotated several quality assessment criteria to an existing discussion corpus. However, the size of the corpus is not large. Annotation to other corpora is an important task. An improvement of the agreement of each criterion will be also an important future research area, although it is essentially a difficult task.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20K12110.

[8] Note that these values cannot be calculated from Table 3 because of a lack of label distribution for each discussion.

References

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-End Argumentation Knowledge Graph Construction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 7367–7374.

Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Transactions on Multimedia*, 18(4):643–658.

Léon Bottou. 1991. Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-Nîmes 91*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Jacob Devlin, Ming-Wei Chang, and and Kristina Toutanova Kenton Lee. 2019. BERT: Pre-training of Deep Bidirectional Ttransformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 4171–4186.

Yuki Hayashi, Fumio Nihei, Yukiko I. Nakano, Hung-Hsuan Huang, and Shogo Okada. 2015. Development of Group Discussion Interaction Corpus and Analysis of the Relationship with Personality Traits. *IPSJ Journal*, 56(4):1217–1227.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6300–6308.

Naoki Mukawa, Tomohiro Nakayama, Hiroko Tokunaga, Junji Yamato, and Naomi Yamashita. 2018. Analysis of Verbal / Nonverbal Expressions of Speakers and Evaluators' Evaluations in Group Discussions. *The IEICE Transactions on Information and Systems*, J101-D(2):284–293.

Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pages 14–20.

Fumio Nihei, Yukiko Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pages 136–143.

Shogo Okada, Yoshihiko Ohtake, Yukiko Nakano, Hayashi Yuki, Hung-Hsung Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating Communication Skills Using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 169–176.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The Discussion Tracker Corpus of Collaborative Argumentation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1033–1043.

Ning Qian. 1999. On the Momentum Term in Gradient Descent Learning Algorithms. *Neural Networks*, 12(1):145 – 151.

T. Shiota and K. Shimada. 2020. The Discussion Corpus toward Argumentation Quality Assessment in Multi-Party Conversation. In *Proceedings of LTLE*, pages 280–283.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 428–437.

Vladimir Vapnik. 2013. *The Nature of Statistical Learning Theory*. Springer science & business media.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 250–255.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 176–187.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 606–615.

Weiqun Xu, Jean Carletta, Jonathan Kilgour, and Vasilis Karaiskos. 2005. Coding Instructions for Topic Segmentation of the AMI Meeting Corpus Version 1.1.

Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. 2016. The Kyutech Corpus and Topic Segmentation Using a Combined Method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR)*, pages 95–104.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 136–141.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–212.

Impact of Distance Measures on Urdu Document Clustering

Zarmeen Nasim
Artificial Intelligence Lab
School of Mathematics and Computer Science
Institute of Business Administration (IBA), Karachi
Znasim@iba.edu.pk

Sajjad Haider
Artificial Intelligence Lab
School of Mathematics and Computer Science
Institute of Business Administration (IBA), Karachi
sahaider@iba.edu.pk

Abstract

Document Clustering aims to group similar documents based on the distance among them. A wide range of distance measures are available in the literature, and selecting an appropriate distance function is a non-trivial task. This paper empirically evaluates four distance measures: Euclidean, Manhattan, Cosine, and Pearson Correlation, on Urdu news headlines. In addition to distance measures, the effect of stemming and lemmatization techniques on clustering is also studied. Unigram-based features and word embedding-based features were used to build a feature matrix. The evaluation results indicate that the frequent unigram features yielded the highest Adjusted Rand Index (ARI) scores on average. Among the four distance measure, the Cosine distance metric was found to be more valuable. Furthermore, the stemming technique was identified to be more useful in contrast to lemmatization for clustering news.

1 Introduction

Document clustering is an unsupervised learning task that aims to group similar documents together and separate dissimilar documents from each other. Most of the clustering algorithms utilize distance functions to identify documents that are syntactically close to each other. The choice of distance measure is critical as it influences the result of clustering.

In literature, clustering and the impact of distance measures have been widely studied on the English language corpus. This paper studies the impact of various distance measures on the clustering of Urdu documents. Urdu is the national language of Pakistan, and despite being spoken by around 170 million people around the globe (Hamdani et al., 2020), it is considered a low-resource language. With the Urdu language support provided on various social media platforms, a huge corpus of user-generated content is now available in digital format. The availability of such enormous data provides opportunities to researchers working in Urdu language processing.

Urdu is a morphologically rich language in contrast to the English language. Several words have various inflections, which adds computational complexity and requires sophisticated models. As a result, the algorithms and techniques developed for the English language cannot be directly applied to the Urdu language corpus due to morphological, syntactical, and lexical differences between both languages. To address the issue of morphological richness of the Urdu language, stemming and lemmatization techniques can be applied to raw text. Therefore, this study also presents the empirical evaluation of stemming and lemmatization on document clustering.

The rest of the paper is organized as follows. Section 2 presents the literature review of document clustering. The methodology is described in Section 3, while Section 4 reports the results of empirical evaluation. Finally, Section 5 concludes the paper.

2 Related Work

This section presents a brief description of the previous research document clustering and the impact of distance functions on clustering.

(Huang, 2008) studied the impact of five different similarity measures, including Euclidean, Cosine, Jaccard, Pearson correlation, and averaged

Kullback-Leibler divergence on partitional clustering. The evaluations were performed on seven different datasets. It was found that the Euclidean distance measure results in the worst performance, whereas the performance of the remaining four measures was similar.

(Aggarwal et al., 2001) studied the impact of distance measures in high dimensional feature spaces. It was theoretically and empirically found that the performance of Lk norm decreases with the increasing value of k in high dimensional spaces. The authors suggested that the Manhattan distance is more appropriate in high dimensional feature space than Euclidean distance.

(Aggarwal et al., 2019) proposed the improvement in the K-means clustering algorithm to deal with the uncertainties in real-world datasets. Further, the authors studied the performance of the proposed clustering algorithm using four distance measures: Euclidean, CityBlock, Cosine, and correlation distance measures. The evaluations were conducted using Davies–Bouldin index and purity metrics. The experiments showed that the correlation distance performed best among other closeness measures as it results in the minimum value of Davies–Bouldin index and maximum purity value.

(A et al., 2013) compared the performance of four different distance measures, including Euclidean, Jaccard, Cosine, and Correlation distance on a clustering task. Purity was used to evaluate the performance of distance measures. It was revealed that Jaccard and correlation distance measures were performing better than Euclidean distance in most cases. (Subhashini and Kumar, 2010) discussed the impact of distance measures on information retrieval and document clustering. They experimented with three distance functions, including Euclidean, Cosine, and Jaccard. Purity metric was used to evaluate the performance of distance measures on clustering tasks. The results showed that the cosine similarity measure and Jaccard index achieved similar performance, whereas the Euclidean distance measure performed worst.

(Bsoul et al., 2014) studied the impact of stemming and lemmatization on Arabic document clustering. Furthermore, the authors also conducted the evaluation of five distance functions for the document clustering task. The distance functions include Cosine, Jaccard, Pearson Correlation, Euclidean, and averaged Kullback-Leibler divergence. The results indicated that the proposed stemming algorithm for the Arabic language yielded good results. Moreover, the experiments also showed that the cosine similarity and Euclidean distance functions achieved the best results compared to other distance measures. (Rahman et al., 2018) studied the effect of various distance measures on Urdu document clustering. The distance measures evaluated in their work included Levenshtein distance, Jaccard index, and Cosine function. The experiments demonstrated that the Jaccard index yielded good results in terms of purity.

In this paper, the impact of stemming and lemmatization is studied for document clustering. The focus of this research is on short-length documents such as News headlines written in the Urdu language. The short-length documents pose the challenge of sparsity in feature space. Furthermore, this paper also identified the best-performing distance measure through empirical evaluation. To the best of our knowledge, such thorough analysis of distance measures, stemming, and lemmatization on short-length document clustering is not performed for the Urdu language.

3 Methodology

This section describes the workflow of experiments conducted to evaluate the impact of distance measures, stemming, and lemmatization.

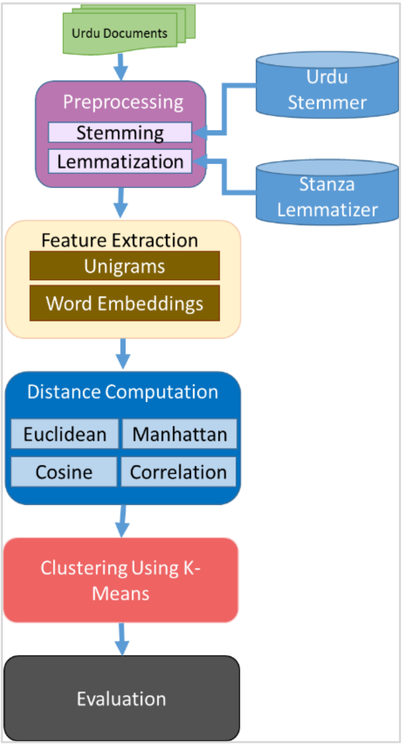


Figure 1: Workflow of Methodology

A pictorial representation of the workflow is presented in Figure 1.

The process involved the following five major steps:

1. Preprocessing
2. Feature extraction
3. Distance computation
4. Clustering
5. Evaluation

Each of these steps is described in the following subsections.

3.1 Preprocessing

The given input corpus is first preprocessed before feature extraction is performed. During preprocessing, the URLs, non-Urdu alphabets, and characters are removed. Punctuation marks and diacritics are also filtered from the input text. Moreover, stopwords are also removed on account of being not crucial for the clustering task. The stop words list, available in the Urdu Hack library[1], is used for the stopwords removal task. After cleaning raw text, stemming and lemmatization are also applied to the cleaned text.

a) **Stemming:** In natural language processing, stemming is the process of transforming a word into its root form. We implemented the stemming approach proposed by (Akram et al., 2009) for Urdu language using the Python language.

b) **Lemmatization:** Lemmatization is the process of grouping all inflections of a word to the base form called *Lemma*. For Urdu lemmatization, the Stanza[2] library is used as it supports the Urdu language along with other numerous languages.

3.2 Feature Extraction

After preprocessing the text, the next step is to convert the text into a feature matrix. For transforming text into a feature matrix, the following two methods are employed.

a) **Unigram Features:** In this method, the text is first tokenized into words. After tokenization, a vocabulary of unique words is built. The length of vocabulary represents the size of the feature matrix. The input text is then transformed into a feature vector. The term frequency-inverse document frequency (TF-IDF) metric is used to weight the feature vector. The matrix built using unigram features represents a sparse matrix where most of the entries are zero.

b) **Word Embeddings:** In recent years, word embeddings have shown tremendous improvements over the bag of words model in various NLP tasks. Word embedding refers to the distributed vector representation of a word in a dense feature space. In this work, pre-trained word embeddings (Kanwal et al., 2019) trained using Word2Vec (Mikolov et al., 2013) algorithm on Urdu news corpus are used. The word embedding model produces a vector representation of a single word. To generate the sentence embeddings, the average of word embeddings is computed.

3.3 Distance Computation

Once the feature matrix is built, a distance function computes the distance among documents. From several distance measures, four functions are used in this paper on account of their popularity. This includes Manhattan, Euclidean, Cosine, and Pearson Correlation distance functions. The details of each of the distance metric are given below:

a) **Manhattan Distance:**

The Manhattan distance, also known as CityBlock, between two data points A (x_1, y_1) and B (x_2, y_2) is the sum of absolute difference. It is computed as:

$$Manhattan\ Distance\ (A,B) = \sum_{i=1}^{N} |A_i - B_i| \tag{1}$$

Where N is the number of dimensions.

b) **Euclidean Distance:**

The Euclidean distance represents the shortest distance between two data points, A (x_1, y_1) and B (x_2, y_2). It is calculated as follows:

$$Euclidean\ Distance\ (A,B) = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \tag{2}$$

c) **Cosine Distance:**

Cosine Similarity measures the cosine of the angle between the two data points A (x_1, y_1) and B (x_2,

[1] https://github.com/urduhack/urduhack

[2] https://stanfordnlp.github.io/stanza/

y_2). The maximum value of similarity represents highly similar documents. This value is subtracted from one to get the distance between two data points. It measures the orientation of the document instead of the magnitude as in Euclidean distance.

The formula given below calculates the cosine distance between A (x_1, y_1) and B (x_2, y_2).

$$Distance\ (A,B) = 1 - \frac{\sum_{i=1}^{N} A_i\ B_i}{\sqrt{\sum_{i=1}^{N} (A_i)^2}\ \sqrt{\sum_{i=1}^{N} (B_i)^2}} \tag{3}$$

d) Pearson Correlation Distance:

Pearson Correlation distance measure is based on the linear correlation between two data points, A (x1, y1) and B (x2, y2). It is limited to only linear associations between the variables. The following formula computes correlation distance.

$$Distance\ (A,B) = 1 - \frac{(A-\bar{A}).(B-\bar{B})}{\|(A-\bar{A})\|_2 \|(B-\bar{B})\|_2} \tag{4}$$

3.4 Clustering

In the previous step, distance measures are used to compute the distance between the documents. In this step, clustering is performed using the K-Means clustering algorithm (Lloyd, 1982). K-Means is a partitional clustering algorithm that assigns documents to different clusters such that the resultant clusters are non-overlapping. The algorithm works as follows:

a) Initialize k centroids randomly

b) Calculate the distance of centroids from each document using the distance function

c) Assign the document to the closest centroid

d) Take the average of the documents to update centroid

e) Reiterate steps (b) – (d) for n number of iterations.

The clustering result produced by the K-Means algorithm depends upon the initial centroids and varies with different seeds. Therefore, all the experiments conducted in this chapter report the average result of five independent runs of the K-Means algorithm. In each run, a different seed value was chosen.

3.5 Evaluation

In this work, the adjusted rand index is used to evaluate and compare the effect of distance measures, stemming, and lemmatization on clustering. Adjusted Rand Index (Hubert and Arabie, 1985) is the measure of similarity between the true cluster labels and the predicted cluster labels. The Rand Index is computed as follows:

$$Rand\ Index = \frac{a+b}{\binom{n}{2}} \tag{5}$$

Where,
n is the number of documents in the clustering,
a refers to the number of documents that are in the same clusters in actual and predicted clustering,
b refers to the number of documents that are in different clusters in actual and predicted clustering.

The adjusted rand index accounts for adjustments due to the number of clusters. It is calculated as:

$$ARI = \frac{RI - Expected\ RI}{\max(RI) - Expected\ RI} \tag{6}$$

The value of ARI is between 0 and 1, where 0 refers to worst quality clusters, and 1 refers to the best quality clusters.

4 Experiments and Results

This section first describes the dataset on which evaluations were performed. Later in this section, we describe the series of experiments that were conducted in this research work.

Cluster Labels	**Count**
کرپشن (Corruption)	491
کرونا وائرس (Coronavirus)	200
سی پیک (CPEC Agreement)	163
نیشنل ٹی ٹونٹی (National T20 Cup)	42
الیکشن کمیشن (Election Commission)	157
جوبائیڈن (Joe Biden)	139
کراچی کے مسائل (Problems of Karachi)	37
موسم (Weather)	355
ڈینگی (Dengue)	166
Total: 1750 Headlines	

Table 1: Dataset Description

Experiment	Features	Euclidean	Cosine	Pearson	Manhattan
Raw Text	All Unigrams	0.708	0.674	0.24	0.544
	Frequent Unigrams	0.7	**0.778**	0.12	0.648
	Word Embeddings	0.41	0.5	0.504	0.424
Stemmed Text	All Unigrams	0.722	0.728	0.308	0.554
	Frequent Unigrams	0.797	**0.825**	0.11	0.617
	Word Embeddings	0.432	0.55	0.522	0.418
Lemmatized Text	All Unigrams	0.668	0.676	0.294	0.618
	Frequent Unigrams	**0.752**	0.742	0.092	0.716
	Word Embeddings	0.522	0.534	0.462	0.482

Table 2: Clustering Results on Urdu News Headlines dataset

4.1 Dataset

The dataset used for empirical evaluation comprised 1750 Urdu news headlines on various topics. The news headlines were fetched from the RSS feeds from the popular Urdu news agencies, including Express[3], UrduPoint[4], Nawae Waqt[5], Voice of America[6] , and BBC Urdu[7]. Table 1 shows basic statistics of the dataset containing news headlines on nine selected keywords. The keywords are used as true cluster labels for the extrinsic evaluation of clustering experiments.

4.2 Clustering Experiments

In this series of experiments, the K-Means clustering algorithm was applied using four different distance measures on raw text, stemmed text and lemmatized text. The raw text refers to the cleaned and preprocessed text without stemming and lemmatization. Three different feature extraction techniques were used to build a feature matrix. The clustering results reported are the average of five independent runs of the K-Means algorithm initialized with different random seeds. K-Means algorithm requires the number of clusters before applying clustering. The value of clusters (k) was set to nine (9) as there were nine topics present in the dataset.

In the first experiment, the raw text was passed to the feature extraction module for extracting various features. K-Means clustering algorithm was then applied on feature matrix for each distance measure. Table 2 presents the results of clustering evaluation on raw text.

It was found that, on average, when frequent unigram features were considered, the highest ARI value was obtained. Furthermore, the results also indicated that the average performance of the clustering algorithm was maximum when cosine distance was used to compute the distance between the news headlines. The Pearson Correlation distance function performed worst on unigram features. However, its performance is almost similar to the cosine distance function on word embedding-based features. This is due to the reason that the mean across word embedding dimensions is zero and the computation of Pearson Correlation distance becomes approximately equal to the cosine distance function.

The second experiment applied stemming to the preprocessed text before the feature extraction stage. Afterward, three different feature extraction techniques were applied, similar to the previous experiment. Clustering was performed on the resultant feature matrix with four distance metrics. As shown in Table 2, on average, frequent unigram

[3] https://www.express.pk/

[4] https://www.urdupoint.com/daily/

[5] https://www.nawaiwaqt.com.pk/

[6] https://www.urduvoa.com/

[7] https://www.bbc.com/urdu

features performed better, as was the case with the first experiment. In addition, it was found that, on average, the clustering algorithm obtained the maximum ARI score when cosine distance was used as a distance function.

In the third experiment, lemmatization was applied to the cleaned text. The results indicated that the frequent unigram features were more effective as they obtained maximum ARI score on average. Moreover, it was found that the clustering algorithm achieved maximum ARI score on average when Euclidean distance was used for distance computation.

To summarize the experimental results presented in Table 2, clustering results were optimal when cosine distance measure was used to compute distance matrix in most experiments. Furthermore, the stemming technique was helpful in contrast to lemmatization as it achieved the highest ARI scores of 0.825, respectively, with frequent unigram features on the dataset.

The aforementioned finding is also supported through the experiments performed by (Bsoul et al., 2014) on Arabic documents. (Bsoul et al., 2014) identified that cosine distance measure produced better clustering in contrast to Euclidean distance measure on clustering task. Furthermore, the authors also highlighted that the stemming obtained better results in comparison to lemmatization on the Arabic document clustering task. Similarly, in our work, stemming yielded optimal ARI score when evaluated on Urdu News headlines clustering.

5 Conclusion

This paper evaluates the impact of stemming and lemmatization on Urdu document clustering. In addition to stemming and lemmatization, the effect of distance measures on clustering short text was also studied. The experiments were performed on the Urdu news headlines corpus. Unigram-based features and word embedding-based features were used to build a feature matrix. The results showed that the frequent unigram features yielded the highest ARI score on average. Among the four distance measure, the Cosine distance metric was more valuable. Furthermore, the stemming technique was identified to be useful in contrast to lemmatization for clustering news headlines. Several extensions to this work are planned for the future. First, we plan to conduct a similar study on Urdu tweets corpus. Second, we would experiment with the recent state-of-the-art feature extraction techniques such as contextualized word embeddings. Lastly, we intend to experiment with various other distance measures to identify the optimal distance metric for the clustering task.

References

Kavitha Karun A, Mintu Philip, and Lubna K. 2013. Comparative Analysis of Similarity Measures in Document Clustering. In *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Swati Aggarwal, Nitika Agarwal, and Monal Jain. 2019. *Performance analysis of uncertain k-means clustering algorithm using different distance metrics*.volume 798. Springer Singapore.

Qurat-ul-Ain Akram, Asma Naseer, and Sarmad Hussain. 2009. Assas-Band , an affix-exception-list based Urdu stemmer . (January):40–46.

Qusay Bsoul, Eiman Al-Shamari, Masnizah Mohd, and Jaffar Atwan. 2014. Distance measures and stemming impact on arabic document clustering. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8870:327–329.

Saboor Hamdani, Rachel Kan, Angel Chan, and Natalia Gagarina. 2020. The Multilingual Assessment Instrument for Narratives (MAIN): Adding Urdu to MAIN. *ZAS Papers in Linguistics*.

Anna Huang. 2008. Similarity measures for text document clustering. In *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019. Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Stuart P. Lloyd. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Atta Rahman, Khairullah Khan, Wahab Khan, Aurangzeb Khan, and Bibi Saqia. 2018. Unsupervised Machine Learning based Documents Clustering in Urdu. *ICST Transactions on Scalable Information Systems*.

R. Subhashini and V. Jawahar Senthil Kumar. 2010. Evaluating the performance of similarity measures used in document clustering and information retrieval. *Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010*:27–31.

Zero-shot and few-shot approaches for tokenization, tagging, and dependency parsing of Tagalog text

Angelina Aquino and **Franz de Leon**
Digital Signal Processing Laboratory, Electrical and Electronics Engineering Institute
University of the Philippines, Diliman, Quezon City, Philippines
{angelina.aquino, franz.de.leon}@eee.upd.edu.ph

Abstract

The grammatical analysis of texts in any written language typically involves a number of basic processing tasks, such as tokenization, morphological tagging, and dependency parsing. State-of-the-art systems can achieve high accuracy on these tasks for languages with large datasets, but yield poor results for languages which have little to no annotated data. To address this issue for the Tagalog language, we investigate the use of alternative language resources for creating task-specific models in the absence of dependency-annotated Tagalog data. We also explore the use of word embeddings and data augmentation to improve performance when only a small amount of annotated Tagalog data is available. We show that these zero-shot and few-shot approaches yield substantial improvements on grammatical analysis of both in-domain and out-of-domain Tagalog text compared to state-of-the-art supervised baselines.

1 Introduction

The grammar of a language is the set of rules and processes which govern the composition of words and sentences in that language. When conducting a grammatical analysis of sentences in a given text, we typically need to perform several tasks such as tokenization, part-of-speech tagging, and syntactic parsing, each of which pertains to a distinct level of grammatical structure. These three levels of analysis are illustrated in Fig. 1 as a dependency tree, one of the most common means of representing a sentence's grammatical structure.

With the advent of machine learning, many natural language processing (NLP) pipelines can now accomplish such analysis tasks with high accuracy for languages with large annotated datasets. However, these systems still yield poor results in a low-resource setting, for languages which have little to no available dependency-annotated data (Zeman et al., 2018). This renders the outputs of such pipelines unsuitable for use in downstream tasks and limits the development of more advanced technologies for these languages.

This is not to say that a language which lacks annotated data is inherently a "low-resource language"—on the contrary, many such languages do in fact possess a wide range of other language resources, including literary works (e.g. books, newspapers, poetry), linguistic references (e.g. dictionaries, grammar textbooks), and the expert knowledge of those proficient with the language. Unlike annotated datasets, these resources cannot be directly fed into existing NLP pipelines as inputs, but it is still possible to encode the information from these resources into usable formats. Given this scenario, we thus ask:

1. Can we create a pipeline for grammatical analysis *without* any dependency-annotated data, using only these alternative language resources?

2. Can we improve the performance of existing pipelines by using alternative resources and methods *together with* a small amount of dependency-annotated data?

In this work, we explore various approaches to achieve these two objectives for the Tagalog language (Section 2), evaluating their performance on educational and news text against state-of-the-art baselines (Section 3). Using the best-performing approaches, we develop a few-

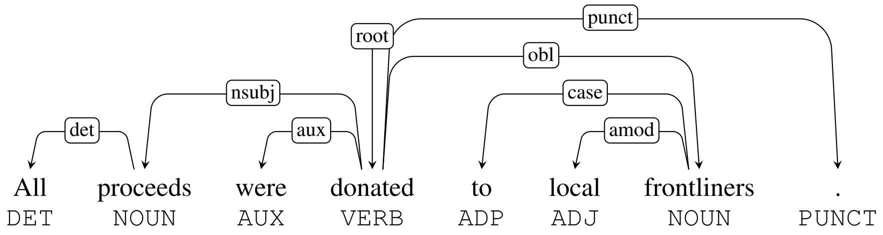


Figure 1: Grammatical analysis (using the Universal Dependencies annotation format[1]) for the sentence: *All proceeds were donated to local frontliners.*

shot pipeline and a zero-shot pipeline (Section 4) which both improve on the current state-of-the-art, and which serve as competent benchmarks for the automated grammatical analysis of Tagalog text. We conduct this work using the Universal Dependencies (UD) framework, which provides cross-linguistically consistent guidelines for grammatical annotation in many languages (Nivre et al., 2016). We situate our work against prior endeavors in automated grammatical analysis for Tagalog, and suggest directions for future research (Section 5). Our scripts, models, and datasets described in this paper are freely available for general use.[2]

2 Candidate approaches

Existing NLP pipelines typically perform several annotation tasks through a chain of individual processing components, each of which can be improved in a low-resource setting using alternative methods and resources. We tested several approaches for implementing the individual pipeline components, then selected the best-performing approaches for each task to create our new pipelines. A description of the specific annotation tasks, and the candidate approaches we have chosen to test for each task, are described in the sections below.

Tokenization is the task of segmenting running text into distinct levels of organization, such as tokens, words, and sentences. We implemented and evaluated the following alternative approach for tokenization:

- *Unsupervised tokenization.* Since tokenization is one of the first preprocessing tasks performed for any text analysis, many tokenizers are included in widely-used NLP libraries. For this task, we used the Python NLTK (Bird, 2006) implementation of the Punkt tokenizer (Kiss & Strunk, 2006), which uses heuristic boundary detection algorithms fine-tuned on the American English *Wall Street Journal* corpus.

Part-of-speech tagging is the task of identifying the parts of speech (e.g. noun, verb, adjective) of words or tokens in a sentence. We implemented and evaluated the following approaches for part-of-speech tagging:

- *Tag conversion.* A POS-tagged corpus for Tagalog was previously developed by Nocon and Borra (2016) using the language-specific MGNN tagset. We developed scripts to map this tagset to the UD POS tagset, and used the corpus text together with the converted tags to train a POS tagger model.

- *Multilingual annotation projection.* In the absence of labeled data for a target language, annotations such as POS tags can be projected from a high-resource source language to a low-resource target language through parallel text. Several massively multi-parallel corpora with Tagalog coverage now exist (Agić et al., 2016; Agić & Vulić, 2019). We adopted the methodology by Plank and Agić (2018) for part-of-speech projection.

Morphological analysis. The UD framework includes levels of annotation for the lemmas (i.e. canonical or uninflected forms) and morphological features (e.g. gender, number, tense) of each word. We implemented and evaluated the following approach for annotating these properties:

- *Rule-based analysis.* Several rule-based methods for morphological analysis of Tagalog were previously implemented (Roxas &

[2] https://universaldependencies.org/guidelines.html

[2] https://github.com/AngelAquino/tagalog-ud-pipeline

Mula, 2008; Cheng, Adlaon, Aquino, Fernandez, & Villanueva, 2017) which have primarily focused on the language's complex verbal morphology. We formed a similar set of rules to identify lemmas and selected UD features, and implemented them as a finite-state transducer.

Dependency parsing is the task of determining the correct head-dependent relationships between words in a given sentence, as well as the labels for each of these relations. We implemented and evaluated the following approaches for dependency parsing:

- *Multilingual annotation projection.* Similar to what was described for POS tagging, annotation projection can also be applied to dependency parsing. Here we implemented the methodology of Agić et al. (2016) for projection of dependency trees.

- *Data augmentation.* When only a small amount of labeled dependency data is available for a given language, several operations can be performed to artificially augment the number of available sentences at training time. In line with the work of Vania et al. (2019), we implemented sentence morphing (i.e. cropping and rotating existing sentences) to produce additional training data.

Neural system with embeddings. The Stanza neural pipeline developed by the Stanford NLP Group has produced state-of-the-art results for dependency parsing for many languages, and for low-resource settings in particular (Qi, Dozat, Zhang, & Manning, 2018). This pipeline consists of a series of neural architectures, and makes use of pre-trained word embeddings (i.e. high-dimensional vector representations encoding the context of words in a given language, trained using large amounts of raw text). We trained Stanza models for tokenization, part-of-speech tagging, morphological analysis, and dependency parsing using the UD-annotated Ugnayan treebank (Aquino & de Leon, 2020) and the fastText word embeddings for Tagalog (Bojanowski, Grave, Joulin, & Mikolov, 2017), and compared the performance of these models to those of the other candidate approaches for the pipeline.

Each of these implementations was evaluated alongside the baseline performance of the UDPipe monolingual Tagalog model, which was determined to be the best supervised system for Tagalog UD parsing among those we evaluated in a previous study (2020). We also used UDify (a state-of-the-art language-agnostic parsing pipeline) as a zero-shot baseline, as a system which achieved competitive performance on Tagalog tagging and parsing without the use of any Tagalog UD annotations in its training.

We tested all approaches using an expanded version of the Ugnayan treebank for Tagalog. The version of this treebank that is currently available on the Universal Dependencies repository[3] consists of 94 sentences (1097 words) of educational text from fiction and nonfiction resources of the DepEd Learning Resource portal. The treebank has been further expanded to include an additional 25 sentences (792 words) of news text from articles by the Philippine Information Agency. We used the educational section of the treebank as training data where required, and the news section purely as testing data.

3 Evaluations

Tokenization. For the task of tokenization, we compare the performance of the UDPipe baseline tokenizer to two alternative systems:

1. The tokenizer component of a Stanza pipeline trained on Ugnayan data as well as fastText Tagalog word embeddings, and

2. the unsupervised Punkt tokenizer, as implemented in Python NLTK.

We report the F_1 scores for each of these models in Table 1. We find that the supervised UDPipe and Stanza tokenizers both outperform the Punkt tokenizer on word and sentence tokenization when tested on in-domain (educational) text, with the UDPipe tokenizer giving marginally better results. On the other hand, the Punkt tokenizer surpasses both UDPipe and Stanza on sentence tokenization of out-of-domain (news) text by a large margin.

Upon inspection of the predicted tokenizations for each system, we find that the Punkt tokenizer gives poor sentence segmentation for quotations (e.g. *"Maraming salamat po!" sabi ko.*) whereas the UDPipe and Stanza tokenizers tend to incorrectly split sentences based on periods following

[3] `https://universaldependencies.org/`

		a. Educational text (raw)			*b. News text (raw)*		
METHOD	TOKEN	WORD	SENT	TOKEN	WORD	SENT	
few-shot	UDPipe (supervised baseline)	**99.27**	**95.67**	**95.41**	91.03	87.81	66.67
	Stanza (supervised + embeds)	99.11	93.74	93.36	93.59	86.97	50.91
zero-shot	Punkt (unsupervised)	97.25	85.08	89.69	**97.10**	**87.58**	**76.92**

Table 1: F_1 scores on tokenization from raw text, tested on both the educational and news sections of the Ugnayan treebank. Results in (a) for both UDPipe and Stanza were averaged over 10-fold cross-validation. **Bold**: highest scores for each test set.

abbreviations (e.g. *Engr.*). These may be attributed to the nature of the different models: the UDPipe and Stanza models were trained on the educational portion of Ugnayan, which does not contain any occurrences of abbreviations, so the model tends to recognize all periods as sentence boundaries, regardless of context. In contrast, the Punkt tokenizer distinguishes periods that follow abbreviations from sentence-terminating periods with high accuracy, but may fail to disambiguate other sentence-terminating punctuation marks which are instead located in the middle of a sentence (as with quotations). These properties may explain the differences in performance between the models on the two datasets: the educational dataset contains 7 sentences with quotations and none with abbreviations (out of 94), while the news dataset contains 4 sentences with abbreviations containing periods (out of 25).

Tagging and morphology. For part-of-speech tagging, we compare the performance of the baseline taggers to four alternatives:

1. The POS tagger component of a Stanza pipeline trained on Ugnayan data as well as fastText Tagalog word embeddings,

2. A POS tagger trained using the ISIP-SAFE Part-of-Speech Tagger corpus, with tags converted from the MGNN tagset into the UD tagset,

3. A POS tagger trained on tags projected from a single source language (English) to Tagalog via alignment of a random selection of parallel English-Tagalog sentences, and

4. A POS tagger trained on tags projected from four source languages (English, Italian, Polish, and Indonesian) to Tagalog via alignment of parallel sentences with optimal coverage across all source-target language pairs.

We used the Bilty bidirectional long-short term memory tagger (Plank, Søgaard, & Goldberg, 2016) to train models 2 to 4. We implemented the tag conversion from the MGNN tagset to the UD tagset as a Python script. We performed this conversion on the ISIP-SAFE Part-of-Speech corpus, and used the corpus text and converted tags to train model 2.

All parallel sentences used in projection for models 3 and 4 come from the JW300 corpus (Agić & Vulić, 2019). For model 3, we selected English for single-source projection as the source language with the highest number of parallel sentences to Tagalog in JW300; we used a randomized selection of text covering approximately 100,000 words (contained in 4831 sentences) to train the model. For model 4, we selected four source languages to represent four different language groups — English for Germanic, Italian for Romance, Polish for Slavic, and Indonesian for Austronesian (to which Tagalog also belongs); here we used the top 5000 Tagalog sentences with optimal alignment coverage averaged across the four source languages.

For models 3 and 4, we trained the source-side taggers using the English-GUM, Italian-ITSD, Polish-PDB, and Indonesian-GSD treebanks from UD v2.6, and used these to tag the source sentences. We then aligned the source and target words using the Eflomal alignment tool (Östling & Tiedemann, 2016), determined the optimal tag for each target word as the tag with the highest sum of tagger confidences across all aligned source words, then projected these optimal tags onto the target words. The target Tagalog words and projected tags were then used as training data for these models.

For morphological feature annotation and lemmatization, we compare the performance of the baselines to three alternatives:

1. The lemmatizer component of a Stanza

		a. Educational text (tokenized)			*b. News text (tokenized)*		
	METHOD	UPOS	FEAT	LEMM	UPOS	FEAT	LEMM
few-shot	UDPipe (supervised baseline)	83.76	94.21	89.79	68.21	**94.19**	75.62
	Stanza (supervised + embeds)	**91.16**	**95.53**	**92.68**	**74.38**	94.07	**79.02**
zero-shot	UDify (zero-shot baseline)	59.80	65.45	71.01	**61.11**	73.48	**72.47**
	POS tag conversion (MGNN)	**68.19**			57.07		
	POS projection (en)	61.17			52.53		
	POS projection (en+id+it+pl)	61.90			57.20		
	Foma FST (v1)		93.53	71.01		**95.33**	70.96
	Foma FST (v2)		**93.71**	**71.19**		**95.33**	70.96

Table 2: F_1 scores on part-of-speech tagging, morphological feature analysis, and lemmatization from tokenized text, tested on both the educational and news sections of the Ugnayan treebank. Results in (a) for both UDPipe and Stanza were averaged over 10-fold cross-validation. **Bold**: highest scores per system type. Gray: highest scores across all models.

pipeline trained on Ugnayan data as well as fastText Tagalog word embeddings,

2. Version 1 of our Foma finite-state transducer for Tagalog morphology, which specifies rules for verbal reduplication and the occurrence of verbal affixes *-um-* and *-in-*, and

3. Version 2 of the Foma FST for Tagalog, which includes all of the rules from Version 1, and adds rules for the occurrence of the verbal affixes *-nag-* and *-hin-*.

We report the F_1 scores for each of these models in Table 2. We find that for part-of-speech tagging and lemmatization, Stanza outperforms both baseline systems on both educational and news texts, while both Stanza and the baseline UDPipe system outperform each of the other alternatives on both datasets by a wide margin. For morphological feature annotation, the Stanza and UDPipe systems also yield marginally better feature annotation than the Foma FST models on the in-domain educational text, but the FST models give a slight improvement over their performance for the out-of-domain news text; all three approaches surpass the zero-shot UDify system on this task by a wide margin. In the absence of UD-annotated Tagalog data, both POS tag conversion and POS projection yield comparable performance to the UDify system, with the tag conversion model and the UDify model performing slightly better on the educational text and news text, respectively.

We can observe that for the part-of-speech tagging task, the use of multiple source languages in the projection approach seems to improve performance over the use of only a single source language; this improvement may be attributed to the coverage-based selection of sentences in the multiple-source model (as opposed to random selection for the single-source model), since a greater amount of coverage has been shown to improve part-of-speech projection performance in prior works (Duong, Cook, Bird, & Pecina, 2013; Plank & Agić, 2018). Further study may be needed to determine the general effects of the number of source languages, as well as their similarity to the target language, on the accuracy of the projection method.

We also note both Versions 1 and 2 of our Tagalog FST yielded the same performance on the news dataset. This indicates that the additional rules included in Version 2 did not correspond to any verbal phenomena that were present in the news dataset, and that in general, the addition of rules may or may not result in an improvement in the morphological analysis of a given text, depending on the contents of that text. Nevertheless, we have shown that even a small number of rules (23 for Version 1, and 27 for Version 2) can account for a large number of verbal phenomena in Tagalog corpora. Moreover, the behavior of such a rule-based system is highly explainable and can be easily expanded to account for any well-defined morphological process (as opposed to the neural models used in the other approaches, which have low explainability, and which can be trained on several examples of some morphological phenomenon without any guarantees that the model is able to generalize to other occurrences of the same phenomenon).

Dependency parsing. For the dependency parsing task, we compare the performance of the baselines to four alternatives:

1. The dependency parser component of a Stanza pipeline trained on Ugnayan data as well as fastText Tagalog word embeddings,
2. The dependency parser component of a UDPipe model trained on Ugnayan data augmented by sentence morphing,
3. The dependency parser component of a UDPipe model trained on dependency trees projected from a single source language (English) to Tagalog via alignment of parallel sentences with optimal coverage, and
4. The dependency parser component of a UDPipe model trained on dependency trees projected from four source languages (English, Italian, Polish, and Indonesian) to Tagalog via alignment of parallel sentences with optimal coverage across all source-target language pairs.

For model 2, we implemented a program which creates additional training sentences through sentence morphing. This program selects all sentences with both `nsubj` and `obj` dependents of the `root` (i.e. subject and direct object clauses of the predicate). It then identifies the word order used in each selected sentence—either `VSO`, `SVO`, or `VOS`, which are the three grammatical word orders in Tagalog—and generates two additional sentences corresponding to the other two word orders by swapping the positions of the clauses in the dependency tree. The educational portion of the Ugnayan treebank was passed as input to this program, and the morphs generated by the program were appended to the original set of sentences, together forming the training data for the parser model.

For models 3 and 4, we used the same set of 5000 Tagalog sentences (and the source sentences parallel to them) as in the multiple-source part-of-speech projection described earlier. We trained source-side parsers using the English-GUM, Italian-ITSD, Polish-PDB, and Indonesian-GSD treebanks from UD v2.6, and used these to generate dependency trees for the source sentences. We then modified the Eflomal alignment tool to print alignment probabilities per word (since the tool originally provided only sentence alignment probabilities in its output), and used this to generate word alignments and probabilities per source-target sentence pair. The parallel sentences and their corresponding alignments were used as inputs to the projection tools developed by Agić et al. (2016) to generate dependency trees projected onto the target sentences through probability weighting and directed maximum spanning tree decoding.

Since these tools only projected dependency edges but not dependency labels, we needed to develop an additional method to predict dependency labels for the projected edges using only available data from the source languages. For this, we trained delexicalized random forest classifiers which predict the dependency label for a word using two features: (1) the POS tag of that word, and (2) the POS tag of the word which heads it. (We say that these models are delexicalized since they only use POS tags as inputs, not the word forms themselves.) We used the same four source-side treebanks as input data for these classifiers. We then took the projected POS tags from our earlier pipeline tests, and used these together with the projected edges in order to predict dependency labels for the target sentences. Finally, we used the trees composed of projected edges, projected POS tags, and predicted dependency labels to train models 3 and 4 above.

We report the F_1 scores for each of these models in Table 3. Among the supervised systems, the UDPipe model trained on augmented data yielded the best UAS and LAS on both datasets, while the Stanza system scored the lowest. Among the zero-shot systems, the UDify model outperformed both projection models by a wide margin. There was a sizeable gap in attachment scores between the supervised and zero-shot models on the educational (in-domain) set; on the other hand, the UDify models yielded scores that approached or even exceeded those of the supervised systems on the news (out-of-domain) set.

We observe that between the two projection models, the single-language model performed slightly better (in terms of LAS on the educational set, and both UAS & LAS on the news set) than the multiple-language model. This runs in contrast to the part-of-speech tagging task, where we found that multiple-source POS projection gave better accuracy than the single-source alternative. Although no study to our knowledge has explored

		a. Educational text (tagged)		*b. News text (tagged)*	
	METHOD	UAS	LAS	UAS	LAS
few-shot	UDPipe (supervised baseline)	75.50	68.95	53.28	43.06
	Stanza (supervised + embeds)	75.08	66.09	51.01	36.99
	data augmentation (morph)	**77.33**	**71.60**	**58.33**	**47.85**
zero-shot	UDify (zero-shot baseline)	**51.96**	**32.18**	**53.28**	**36.62**
	projection (en)	26.62	21.33	26.77	18.81
	projection (en+id+it+pl)	28.08	20.69	24.24	14.77

Table 3: F_1 scores on dependency parsing from gold-tagged text, tested on both the educational and news sections of the Ugnayan treebank. Results in (a) for the UDPipe baseline, Stanza, and UDPipe data augmentation models were averaged over 10-fold cross-validation. **Bold**: highest scores per system type. Gray: highest scores across all models.

the relationship between the number of source languages and the accuracy of the resulting dependency projection, prior works have shown that the best projection approach, and the best source languages, vary across different target languages (Johannsen, Agić, & Søgaard, 2016; Agić et al., 2016). Further work may be needed to determine the effect of each of these parameters on projection for Tagalog and for other Philippine languages.

Beyond the results shown here, we have also observed that although the Stanza model yielded slightly lower average scores than the baseline UDPipe model when parsing from gold-tagged text, the Stanza model outperforms the UDPipe baseline on average when parsing using system-predicted part-of-speech tags and morphological annotations. We initially found this result counter-intuitive, since both parsers were configured to use system-predicted tags during their training, and we have shown in the previous section that the Stanza system produces more accurate tags (i.e. tags that are closer overall to the gold tags) than the UDPipe parser. Hence, we expected that the Stanza parser similarly outperforms the UDPipe parser on gold-tagged data. However, once we inspected the scores of both systems on each of the 10 cross-validation folds for the education set, we found that Stanza in fact gave better UAS for 5 out of 10 folds, and equal or better LAS for 4 out of 10 folds. (In comparison, the augmented model gives better or equal UAS than the UDPipe baseline for 8 out of 10 folds, and better or equal LAS for 9 out of 10 folds.) We therefore conclude that the Stanza and UDPipe baseline models give comparatively equal parsing performance, whereas data augmentation tends to improve on both of these systems overall.

4 Final pipelines

For our final pipelines, we differentiate between two low-resource data settings: few-shot (i.e. when a small amount of UD-annotated training data is available), and zero-shot (i.e. when no UD-annotated data is available). Based on our experiments above, we propose the use of the following pipelines in each setting:

1. *Few-shot.* The Stanza pipeline, trained with word embeddings and UD-annotated data augmented through sentence morphing

2. *Zero-shot.* A pipeline consisting of: the Punkt unsupervised tokenizer; the Bilty POS tagger, trained on tags converted from existing resources for the target language (e.g. a dictionary or a tagged corpus); the Foma morphological analyzer, compiled with rules written based on existing grammar texts for the target language; and the UDify universal parser model

For our final evaluation, we compare the performance of these two pipelines to the UDPipe Indonesian-GSD model, which was the cross-lingual model with the best LAS among those we evaluated in a previous study (2020).

We report the F_1 scores for each of these models in Table 4. When testing on educational text, we find that the few-shot Stanza pipeline (trained in-domain) outperforms both the zero-shot and cross-lingual pipelines on all tokenization, tagging, and parsing tasks. The results on the news data, on the other hand, are more varied:

- For token- and word-level tokenization, the zero-shot pipeline gives the highest average scores, performing slightly better than

a. Educational text (raw)	*Tokenization*			*Tagging*			*Parsing*	
PIPELINE	TOKEN	WORD	SENT	UPOS	FEAT	LEMM	UAS	LAS
few-shot	**98.75**	**93.56**	**98.57**	**85.12**	**90.55**	**87.21**	**67.29**	**60.15**
zero-shot	97.25	85.08	89.69	59.57	78.75	66.57	48.53	33.29
cross-lingual	97.40	85.22	92.38	27.95	46.56	65.92	18.78	9.41

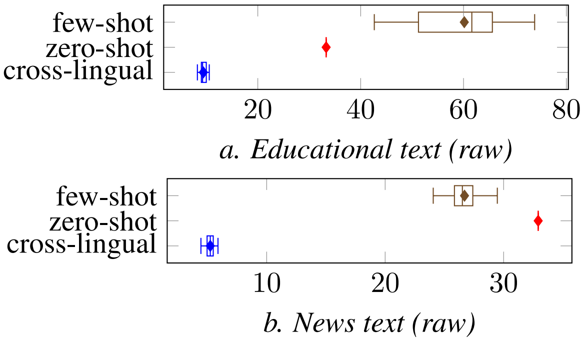
b. News text (raw)	*Tokenization*			*Tagging*			*Parsing*	
PIPELINE	TOKEN	WORD	SENT	UPOS	FEAT	LEMM	UAS	LAS
few-shot	90.85	84.15	32.14	**61.51**	78.45	65.98	35.26	26.70
zero-shot	**97.29**	**87.74**	83.02	53.77	**83.80**	66.89	**44.07**	**32.92**
cross-lingual	96.64	87.13	**86.89**	30.77	55.05	**67.29**	10.12	5.24

Table 4: F_1 scores on tokenization, tagging, and parsing from raw text, tested on both the educational and news sections of the Ugnayan treebank. **Bold**: highest scores for each test set.

the cross-lingual pipeline, with the few-shot pipeline lagging behind on token-level segmentation.

- For sentence-level tokenization, the cross-lingual pipeline gives the best performance, followed closely by the zero-shot pipeline, with both systems far surpassing the few-shot pipeline (trained out-of-domain).
- For part-of-speech tagging and morphological feature annotation, the gaps between scores of each of the systems are more pronounced. The few-shot pipeline does best on the POS task, the zero-shot pipeline does best on feature annotation, and the cross-lingual pipeline is well behind the other two on both tasks.
- For lemmatization, the cross-lingual pipeline achieves the best average, but the other two systems give roughly similar performance on the task.
- For dependency parsing, the zero-shot pipeline outperforms the few-shot pipeline on both unlabeled and labeled attachment, with both systems outperforming the cross-lingual pipeline by a large margin.

Since our focus is on dependency parsing, we use the labeled attachment score as the primary measure of system performance. For this, we compare the ranges of labeled attachment scores achieved by each model across 10 folds of cross-validation, wherein the training and dev partitions are partitioned differently for each fold. (None of the zero-shot components require train/dev partitions in their development, so only a single model is used per component across all tests, and only a single score is reported for each test set.) These results are visualized in Fig. 2.



a. Educational text (raw)

b. News text (raw)

Figure 2: LAS spreads for 10-fold cross-validation on tokenization, tagging, and parsing from raw text, tested on the educational section of the Ugnayan treebank.

We find that when testing in-domain (on educational text), the few-shot pipeline yields higher scores across all folds than the zero-shot pipeline. The opposite is true when testing out-of-domain (on news text): here, the zero-shot pipeline outperforms the few-shot pipeline on all folds. We can see that the performance of the zero-shot system stays consistent across both types of data (with LAS between 32% and 34% for both). In contrast, the performance of the few-shot system is highly sensitive to the train-dev split used at training time (with a difference of over 30% LAS between the lowest and highest scores achieved on the educational text, and to the type of data encountered at training versus testing (with the median LAS dropping by over 30% from the in-domain to the out-of-domain scenario). We therefore recommend the use of the zero-shot pipeline for general Tagalog parsing purposes, and would only consider using the few-shot pipeline if the data to be parsed is in the same domain as any annotated datasets available for training the pipeline.

In addition, we find that both the few-shot and zero-shot pipelines outperform all instances of the cross-lingual pipeline by 18% or more LAS. This indicates that these new pipelines improve on Tagalog parsing performance versus state-of-the-art pipelines trained on high-resource cross-lingual data (in this case, UDPipe trained on an Indonesian treebank containing over $100,000$ tokens, which yielded better LAS than several other cross-lingual alternatives); we have shown that this can be achieved through the use of alternative Tagalog language resources in both a few-shot setting (with less than $1,000$ tokens of UD-annotated training data) and a zero-shot setting (using no UD-annotated target-language data). Sample outputs for all three pipelines on news text are rendered in graphical form in Fig. 3.

5 Discussion

Tagalog is an Austronesian language spoken as a first language by around a quarter of the 100 million total population of the Philippines, and as a second language by the majority of the country (Eberhard, Simons, & Fennig, 2022). Its standardized form, Filipino, is one of two official languages for communication and instruction in the Philippines (the other being English). It has an abundance of language resources available to the public, including academic texts, creative literature, newspapers, radio & TV broadcasts, and social media content. Furthermore, it has long been a language of interest for linguists worldwide due to its complex verbal morphology and distinctive voice system.

Although Tagalog has both a thriving literary tradition and an established body of linguistic research behind it, computational NLP for Tagalog is only an emerging field of study, with much progress yet to be made compared to the other major languages of the world. Some of the first systems for automated grammatical analysis of Tagalog include the rule-based morphological analyzer by Fortes (2002), the template-based POS tagger by Rabo (2004) whose tagset was the basis for the later MGNN tagset used in our tag conversion approach, and the graph-based dependency parser by Manguilimotan & Matsumoto (2011). More recent statistical and neural implementations have been tested by Go & Nocon (2017) for POS tagging, and by Yambao & Cheng (2020) for morphological analysis, showing substantial improvements for each task over previous works.

One of the shortcomings of the systems above is that their output annotations are non-standard and are not directly comparable to the results achieved for other languages. The first cross-linguistically compatible Tagalog treebank was created by Samson (2018) under the UD standard. We subsequently conducted, to our knowledge, the first comparative evaluation of supervised UD parsing for Tagalog (2020) using the TRG treebank by Samson as well as our own Ugnayan treebank.

Our experiments in this paper have shown that the use of various available language resources through simple data conversion and generation methods can improve performance over supervised grammatical analysis with annotated data alone. The approaches we have tested were chosen primarily on the basis of:

- the language resources that were available to us for Tagalog;
- the limits in computational capacity of the consumer laptop used throughout this research; and
- the grammatical properties of the Tagalog language.

We hypothesize that the applicability of these methods to other languages will in turn be dependent on the above factors. For example, the Tagalog FST models we created for morphological analysis may be easily modified to suit other Philippine languages like Bikolano and Ilokano with similar verb infixation patterns, but would not be useful for more analytic languages like Vietnamese, and may require significant expansion for more polysynthetic languages like Kunwinjku, as seen in Lane & Bird (2019). The dependency projection method for zero-shot parsing was not effective here for Tagalog, but it may yield better results for target languages with high similarity to annotation-rich sources (e.g. Germanic and Romance languages) or if many more source languages and higher compute are used, as in Agić et al (2016). Language resources beyond those used in this paper, such as speech corpora and semantic networks, may require different encoding tech-

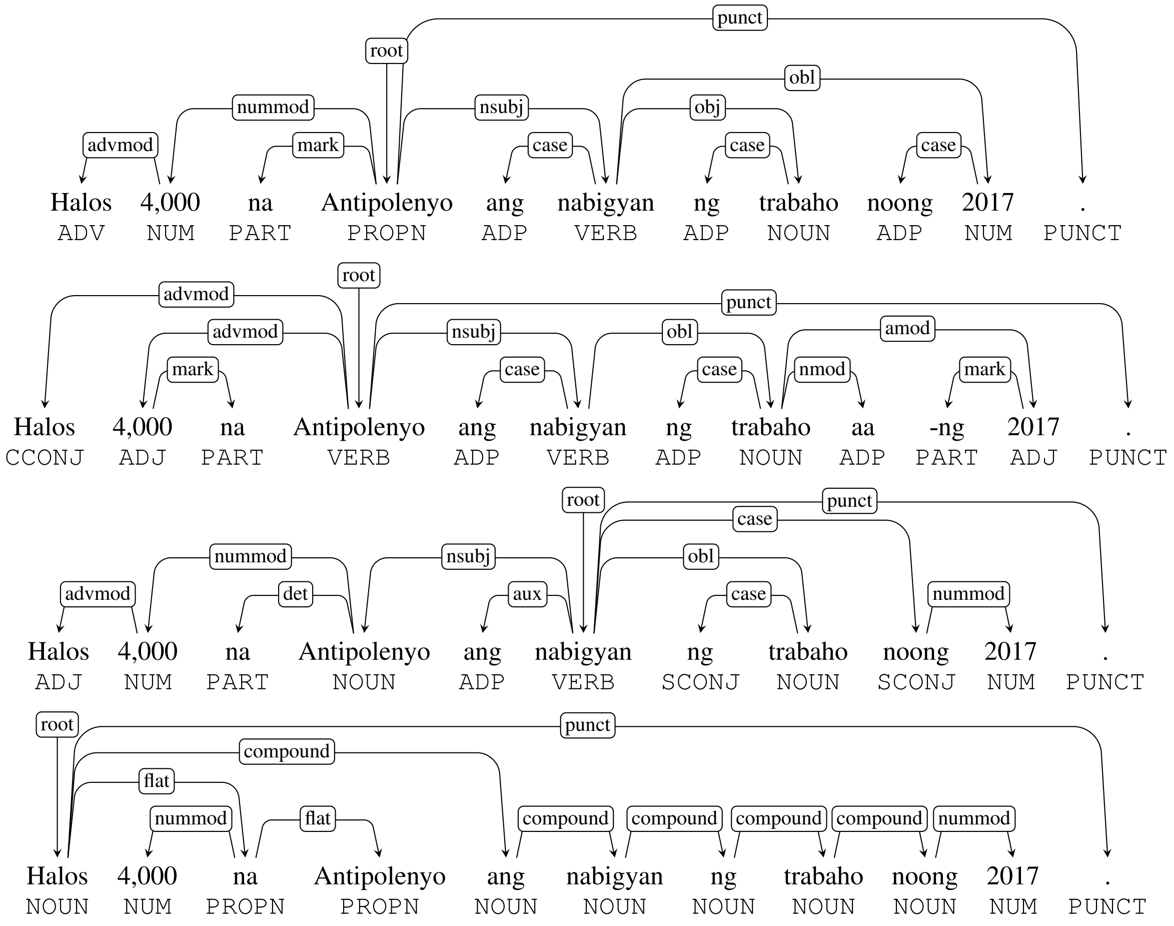

punct
root
obl
nummod
nsubj
obj
advmod
mark
case
case
case
Halos 4,000 na Antipolenyo ang nabigyan ng trabaho noong 2017 .
ADV NUM PART PROPN ADP VERB ADP NOUN ADP NUM PUNCT
root
advmod
punct
advmod
nsubj
obl
amod
mark
case
case
nmod
mark
Halos 4,000 na Antipolenyo ang nabigyan ng trabaho aa -ng 2017 .
CCONJ ADJ PART VERB ADP VERB ADP NOUN ADP PART ADJ PUNCT
root
punct
case
nummod
nsubj
obl
advmod
det
aux
case
nummod
Halos 4,000 na Antipolenyo ang nabigyan ng trabaho noong 2017 .
ADJ NUM PART NOUN ADP VERB SCONJ NOUN SCONJ NUM PUNCT
root
punct
compound
flat
nummod
flat
compound
compound
compound
compound
nummod
Halos 4,000 na Antipolenyo ang nabigyan ng trabaho noong 2017 .
NOUN NUM PROPN PROPN NOUN NOUN NOUN NOUN NOUN NUM PUNCT

Figure 3: Ugnayan treebank annotation (1st row) and outputs of the few-shot, zero-shot, and cross-lingual pipelines (2nd to 4th rows) for the following sentence: *Halos 4,000 na Antipolenyo ang nabigyan ng trabaho noong 2017* (Around 4,000 Antipolenyos were granted jobs in 2017).

niques and system architectures to be useful for the analysis tasks investigated here.

6 Conclusion

In this work, we have explored methods for leveraging alternative Tagalog language resources to improve few-shot parser performance. We have also shown that in the absence of any annotated Tagalog data, a pipeline of zero-shot methods for tokenization, tagging, and parsing also yields better results than cross-lingual models, and can even outperform the improved monolingual models when testing on out-of-domain text.

We recommend the use of the zero-shot pipeline—consisting of unsupervised tokenization, part-of-speech tag conversion, finite-state morphological analysis, and a multilingual BERT-based parser—for general-purpose, automated grammatical analysis of Tagalog text, as it gives consistent performance for all UD annotation tasks across the two domains we have investigated here. If the text to be analyzed is sufficiently similar to the available Tagalog training data, we also recommend the use of supervised monolingual modeling, as it has been shown to yield substantial improvements over the other alternatives tested here when annotating in-domain text.

There exist many possible avenues for building on this work, including but not limited to the creation of new treebanks and the expansion of existing treebanks, the use of domain adaptation methods for out-of-domain parsing, the exploration of Tagalog as a pivot between foreign and Philippine languages, and the investigation of methods for annotating and parsing code-switched (e.g. mixed Tagalog and English) text. We hope that our initial steps here will encourage further work towards robust grammatical analysis for Tagalog, and for many other local languages here and abroad.

Acknowledgments

We are grateful to Professors Mary Ann Bacolod, Anastacia Alvarez, and Rhandley Cajote for their guidance and support over the course of this project. We also thank the four anonymous reviewers for their insightful comments on an earlier draft of this paper. This work was funded by the U.P. Teaching Assistantship Program.

References

Agić, Ž., Johannsen, A., Plank, B., Martínez Alonso, H., Schluter, N., & Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, *4*, 301–312. Retrieved from `https://www.aclweb.org/anthology/Q16-1022` doi: 10.1162/tacl_a_00100

Agić, Ž., & Vulić, I. (2019, July). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3204–3210). Florence, Italy: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P19-1310` doi: 10.18653/v1/P19-1310

Aquino, A., & de Leon, F. (2020, December). Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog. In *Proceedings of the fourth workshop on universal dependencies (udw 2020)* (pp. 8–15). Barcelona, Spain (Online): Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.udw-1.2`

Bird, S. (2006, July). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 69–72). Sydney, Australia: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P06-4018` doi: 10.3115/1225403.1225421

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from `https://www.aclweb.org/anthology/Q17-1010` doi: 10.1162/tacl_a_00051

Cheng, C., Adlaon, K. M., Aquino, M., Fernandez, E., & Villanueva, K. (2017, March). MAG-Tagalog: A rule-based Tagalog morphological analyzer and generator. In *Proceedings of the 17th philippine computing science congress* (pp. 171–178). University of San Carlos, Talamban Campus, Cebu City, Philippines: Computing Society of the Philippines.

Duong, L., Cook, P., Bird, S., & Pecina, P. (2013, August). Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 634–639). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P13-2112`

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2022). *Ethnologue: Languages of the world* (25th ed.). Dallas, Texas: SIL International. Retrieved from `http://www.ethnologue.com`

Fortes, F. C. L. (2002). *A constraint-based morphological analyzer for concatenative and non-concatenative morphology of Tagalog verbs* (Master's thesis). De La Salle University, Manila, Philippines.

Go, M. P., & Nocon, N. (2017, November). Using Stanford part-of-speech tagger for the morphologically-rich Filipino language. In *Proceedings of the 31st pacific asia conference on language, information and computation* (pp. 81–88). Manila, Philippines: The National University (Phillippines). Retrieved from `https://www.aclweb.org/anthology/Y17-1014`

Johannsen, A., Agić, Ž., & Søgaard, A. (2016, August). Joint part-of-speech and dependency projection from multiple sources. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 561–566). Berlin, Germany: Association for Computational Linguistics. Retrieved from `https://www.aclweb`

.org/anthology/P16-2091 doi: 10.18653/v1/P16-2091

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, *32*(4), 485–525. Retrieved from https://www.aclweb.org/anthology/J06-4003 doi: 10.1162/coli.2006.32.4.485

Lane, W., & Bird, S. (2019, 4–6 December). Towards a robust morphological analyzer for kunwinjku. In *Proceedings of the the 17th annual workshop of the australasian language technology association* (pp. 1–9). Sydney, Australia: Australasian Language Technology Association. Retrieved from https://www.aclweb.org/anthology/U19-1001

Manguilimotan, E., & Matsumoto, Y. (2011, December). Dependency-based analysis for Tagalog sentences. In *Proceedings of the 25th pacific asia conference on language, information and computation* (pp. 343–352). Singapore: Institute of Digital Enhancement of Cognitive Processing, Waseda University. Retrieved from https://www.aclweb.org/anthology/Y11-1036

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., ... Zeman, D. (2016, May). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1659–1666). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L16-1262

Nocon, N., & Borra, A. (2016, October). SMTPOST using statistical machine translation approach in Filipino part-of-speech tagging. In *Proceedings of the 30th pacific asia conference on language, information and computation: Posters* (pp. 391–396). Seoul, South Korea: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/Y16-3010

Östling, R., & Tiedemann, J. (2016). Efficient word alignment with Markov chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, *106*, 125–146. doi: 10.1515/pralin-2016-0013

Plank, B., & Agić, Ž. (2018, October-November). Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 614–620). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D18-1061 doi: 10.18653/v1/D18-1061

Plank, B., Søgaard, A., & Goldberg, Y. (2016, August). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 412–418). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-2067 doi: 10.18653/v1/P16-2067

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018, October). Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 160–170). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K18-2016 doi: 10.18653/v1/K18-2016

Rabo, V. S. (2004). *TPOST: A template-based, n-gram part-of-speech tagger for Tagalog* (Master's thesis). De La Salle University, Manila, Philippines.

Roxas, R., & Mula, G. (2008, November). A morphological analyzer for Filipino verbs. In *Proceedings of the 22nd pacific asia conference on language, information and computation* (pp. 467–473). The University of the Philippines Visayas Cebu College, Cebu City, Philippines: De La Salle University, Manila, Philippines. Retrieved from https://www.aclweb.org/anthology/Y08-1050

Samson, S. D. (2018). *A treebank prototype of Tagalog* (Undergraduate thesis). University of Tübingen, Germany.

Vania, C., Kementchedjhieva, Y., Søgaard, A., & Lopez, A. (2019, November). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1105–1116). Hong Kong, China: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D19-1102` doi: 10.18653/v1/D19-1102

Yambao, A., & Cheng, C. (2020, December). Feedforward approach to sequential morphological analysis in the Tagalog language. In *2020 international conference on asian language processing* (pp. 81–85). Monash University, Kuala Lumpur, Malaysia: IEEE.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., … Petrov, S. (2018, October). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 1–21). Brussels, Belgium: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/K18-2001` doi: 10.18653/v1/K18-2001

Syntactic Analysis of *-kah* in Malay Polar Questions

Serena P. W. Jong
Chinese University of Hong Kong
Shatin, Hong Kong SAR
serenajpwun@gmail.com

Lawrence Y. L. Cheung
Chinese University of Hong Kong
Shatin, Hong Kong SAR
yllcheung@cuhk.edu.hk

Abstract

This paper talks about the question particle *-kah*, that is commonly used in both constituent and polar questions in Malay. This particle is interesting because it can appear in different positions in polar questions. Despite previous research on the topic, there is a lack of comprehensive description and syntactic analysis of *-kah* in Malay polar questions. We found that *-kah* occupies four major syntactic positions: (i) after the sentence-initial verb/modal, (ii) after the sentence-initial phrase, (iii) after the entire sentence, and (iv) after the sentence-initial auxiliary *ada*. These different syntactic positions are attributable to three syntactic processes, i.e. head movement, phrasal movement and base-generated *ada*. This paper suggests that the diverse syntactic positions can receive a unified account by assuming that *-kah* is a bound C head that requires phonological support like the *-ed* in English using any of the three syntactic mechanisms.

1 Overview

In Malay, the particle *-kah* is typically used in constituent questions and polar questions, as in (1) and (2) respectively. It has been commonly referred to as the question particle in the literature. While some studies on Malay constituent questions can be found, not much has been said about the particle *-kah* in Malay polar questions.

(1) Siapa-kah nama kamu?
who-Q name you
'What is your name?'

(2) Besar-kah rumah Ali?
big-Q house Ali
'Is Ali's house big?'

At first glance, the distribution of particle *-kah* in polar questions appears to be somewhat arbitrary. As shown in (1) and (3b), *-kah* can show up after a word near the beginning of the polar question. In (3a), however, *-kah* occurs in the sentence-final position.

(3) a. Ali pergi ke hospital-kah?
Ali go to hospital
'Did Ali go TO THE HOSPITAL?'

b. Pergi-kah Ali ke hospital?
go-Q Ali to hospital
'Did Ali go to the hospital?'

Where can *-kah* occur in Malay polar questions? Are these instances of *-kah* related? In this paper, we attempt to answer these questions by addressing the strategies used in all types of Malay polar questions involving *-kah* and propose a unified analysis to account for the particle *-kah*. We show that a single economy principle is sufficient in capturing the large variety of options for polar question formation involving *-kah*.

The rest of the paper is structured as follows. In Section 2, we will present a short review of the literature related to Malay polar questions. We show that previous studies are still not comprehensive enough in the description of *-kah* in Malay polar questions. In Section 3, we provide some of the facts in Malay to facilitate our discussion in the following sections and also briefly talk about the various types of Malay polar questions. Section 4 will propose a unified analysis in which multiple strategies can be used in Malay polar questions to support *-kah*. Finally, a summary of the whole paper will be included in Section 5.

2 Previous descriptions of Malay polar questions

Previous research on Malay polar questions is rather limited. They focus on describing the different types of polar questions in the language. Karim et al. (2008) have identified some of the patterns allowed in Malay polar question formation, i.e. SVO pattern with rising intonation at the end of the sentence[1] (4), SVO pattern with *-kah* appearing at the end of the sentence (5) and phrases/words appearing before *-kah* near the beginning of the sentence (6-7).

(4) Kamal me-mandu lori?[2]
Kamal ACT-drive lorry
'Does/Can Kamal drive a lorry?'
(Karim et al., 2008, p.442)

(5) Mereka berlepas pagi tadi-kah?
they leave morning just.now-Q
'Did they leave THIS MORNING?'
(Karim et al., 2008, p.442)

(6) Pagi tadi-kah mereka berlepas?
morning just.now-Q they leave
'Was it this morning that they left?'
(Karim et al., 2008, p.443)

(7) Cantik-kah istana raja itu?
beautiful-Q palace king that
'Is that palace beautiful?'
(Karim et al., 2008, p.443)

While the Karim et al.'s (2008) description of the different types of Malay polar questions serves as an important starting point for the study of Malay polar question formation, the syntactic analysis of the distribution pattern of *-kah* is lacking. Most importantly, they did not say why *-kah* in (5)—(7) can occupy different positions.

Fortin (2009) described three strategies for forming polar questions in Indonesian. As Malay and Indonesian are very close to each other, the study is useful to our understanding of Malay polar questions. The first strategy only involves question intonation (8); the second strategy involves a question particle, *apa(kah)* (9); and the third strategy involves a fronted constituent marked with *–kah* (10).

(8) Siti sudah pulang?
Siti TEMP go.home
'Did Siti go home?'
(Fortin, 2009, p.32)

(9) Apa(kah) Siti sudah pulang?
what(Q) Siti TEMP go.home
'Did Siti go home?'
(Fortin, 2009, p.32)

(10) [Sudah pulang]+kah Siti?
TEMP go.home]+Q Siti
'Did Siti ALREADY GO HOME?'
(Fortin, 2009, p.32)

Fortin (2009) has proposed a unified analysis and argued that Indonesian questions are focus constructions as the particle *-kah*, which functions as a focus marker is present. Further evidence supporting this claim has been provided, including the similarities between *-kah* and the non-interrogative focus marker *-lah* in Indonesian, as illustrated below.

(11) a. [Pintu itu]+lah *(yang) Ali (*mem-)buka.
[door that]+FOC COMP Ali (*ACT-)open
'It is that door that Ali opened.'
(Fortin, 2009, p.36)

b. [Pintu yang mana]+kah *(yang) Ali (*mem-)buka?
[door COMP which]+Q COMP Ali (*ACT-)open
'Which door did Ali open?'

Although this may seem like a desirable analysis, Fortin (2009) has noted that this analysis does not necessarily extend to Malay

[1] This pattern will not be discussed in this paper as it is not associated with *-kah*.

[2] All the examples cited from Karim et al. (2008) do not come with English translation. Whenever examples are cited from Karim et al. (2008), English translation will be provided.

as both languages may differ in some respects. For one thing, there are some differences in the strategies used in polar question formation between the two languages. Besides, not much has been said about the distribution of *-kah*, in particular, in Malay polar questions. Therefore, this will be further discussed later in this paper.

3 On the basics of Malay polar questions

3.1 Basic word order of Malay

As a starting point for the analysis of Malay polar question formation involving *-kah*, we will briefly talk about the basic word order of Malay. Malay is a head-initial language and has an SVO word order. As shown in (12a), the subject *Ali* is followed by the verb *membaca* and the verb is followed by the object *buku*. When a modal is present in the sentence, the modal will always precede the verb, as shown in (12b).

(12) a. Ali mem-baca buku itu.
Ali ACT-read book that
'Ali read that book.'
(Karim et al., 2008, p.350)

b. Ali harus mem-baca buku itu.
Ali must ACT-read book that
'Ali must read that book.'

3.2 Distribution of *-kah* in Malay polar questions

In this section, we will provide a preliminary sketch of the types of polar questions involving *-kah* in Malay. In general, there are four types of polar questions with *-kah* attached in Malay:

1) Type 1: polar question with *-kah* appearing after a verb (13b)/modal (14c)
2) Type 2a[3]: polar question with *-kah* appearing after a phrase (15b)
3) Type 2b: polar question with *-kah* appearing after an entire sentence with SVO word order (15c)
4) Type 3: polar question with *-kah* appearing after the auxiliary *ada*[4] (16b)

Type 1: As shown in (13b), the verb *makan* appears before *-kah* and the subject *Ali*. This is different from the corresponding declarative sentence in (13a), where the verb normally appears after the subject. Note that this type of polar question is allowed in both formal and colloquial Malay.

(13) a. Ali makan kek itu.
Ali eat cake that
'Ali ate that cake.'

b. Makan-kah Ali kek itu?
eat-Q Ali cake that
'Did Ali eat that cake?'

If a modal is present in the sentence, the verb can no longer be placed before *-kah*. As illustrated in (14b), moving the verb *makan* to a position before *-kah* will result in an ungrammatical sentence as the modal *boleh* 'can' is present in the sentence. Conversely, the modal *boleh* can appear before *-kah*, as shown in (14c).

(14) a. Ali boleh makan kek itu.
Ali can eat cake that
'Ali can eat that cake.'

b *Makan-kah Ali boleh kek itu?
eat-Q Ali can cake that
'Intended: Can Ali eat that cake?'

c. Boleh-kah Ali makan kek itu?
can-Q Ali eat cake that
'Can Ali eat that cake?'

Type 2a: In (15b), it can be observed that the locative phrase *ke sekolah* can appear before -*kah*. Similar to Type 1 polar questions, Type

[3] Note that 2 and 3 are labelled as Type 2a and 2b respectively as we will adopt an analysis that will collapse the two as one in the following section.

[4] Not much has been said about *ada* in Malay grammar reference books. However, *ada* appearing in polar questions behaves similarly like the dummy auxiliary *do* in English. Thus, we will refer *ada* as an auxiliary in this paper.

2a polar questions are commonly used in both formal and colloquial Malay.

Type 2b: As observed in (15c), *-kah* appears at the end of the whole sentence with SVO word order. It should be mentioned, however, this type of polar question is only allowed in colloquial Malay.

(15) a. Dia ke sekolah.
he/she to school
'He/She went to school.'
(Karim et al., 2008, p.442)

b. Ke sekolah-kah dia?
To school-Q he/she
'Was it to school that he/she went?'
(Karim et al., 2008, p.443)

c. Dia ke sekolah-kah?
he/she to school-Q
'Is he/she GOING TO SCHOOL?'
(Karim et al., 2008, p.442)

Type 3: As illustrated in (16b), *-kah* appears after *ada* in this type of polar question. Note that for this type of polar question, *ada* is not found in the corresponding declarative sentence in (16a) but is inserted to the polar question. It resembles the insertion of dummy "do" in English polar questions.

(16) a. Siti me-nelefon Ali semalam.
Siti ACT-call Ali yesterday
'Siti called Ali yesterday.'

b. Ada-kah Siti me-nelefon Ali
ada-Q Siti ACT-call Ali
semalam?
yesterday
'Did Siti call Ali yesterday?'

3.3 More on the responses to Malay polar questions

The polar questions in Malay are also known as *ayat tanya tertutup*, meaning closed questions (Karim et al., 2008). There are two possible ways of responding to polar questions in Malay, i.e. *ya* to indicate positive response (17b) and *bukan/tidak* to indicate negative response (17c) and (18b).

(17) a. Minum-kah Ali susu itu?
drink-Q Ali milk that
'Did Ali drink that milk?'

b. Ya, Ali minum susu itu.
yes Ali drink milk that
'Yes, Ali drank that milk.'

c. Tidak, Ali tidak minum
no Ali no drink
susu itu.
milk that
'No, Ali didn't drink that milk.'

(18) a. Ali-
kah yang minum
Ali-Q that drink
susu itu?
milk that
'Was it Ali who drank that milk?'

b. Bukan, yang minum susu
no that drink milk
itu bukan Ali.
that no Ali
'No, it was not Ali who drank that milk.'

4 Unified analysis of *-kah*

Based on the examples provided in Section 3, it seems that particle *-kah* can appear in different positions in Malay polar questions. As far as we know, no prior work has been done to explain why *-kah* can occur in different positions in Malay polar questions. In this section, we will provide a unified analysis of *-kah* in Malay polar questions. Essentially, it is argued that *-kah* is a bound suffix that must attach to a host element such as a word or a phrase. To justify this, we will begin by showing the different types of Malay polar questions mentioned in Section 3, in fact, are different strategies used to support *-kah* for phonological spellout. In general, three strategies have been identified to provide a host element for *-kah* in Malay polar questions:

1) head movement
2) phrasal movement
3) base-generated *ada*

4.1 Head movement

Verb movement in Malay polar questions
As shown in (13b), it is possible for the verb to move to sentence-initial position when forming a polar question in Malay. However, this is only

possible when the verb is fronted to combine with *-kah*. As shown in (19b), only the verb *minum* (but not the entire verb phrase *minum air itu*) is moved to sentence-initial position, while the rest of the phrase *air itu* remains in its original position. As only the V head is moved in Type 1 polar questions, we analyze this as a head movement. For such movement, we assume that *minum* which originates from the V position, moves to the T position and then moves from T to the C position, as shown in Figure 1. The movement is consistent with the Head Movement Constraint (HMC), which assumes that a head generally can only move upward to adjoin to the next c-commanding head (Travis, 1984). The analysis is similar to French, where the main verb is allowed to be moved to the front of the sentence (Freidin, 2012). In (20b), only the main verb moves to T, and then from T to C.

(19) a. Siti minum air itu.
Siti drink water that
'Siti drank that water.'

b. Minum-kah Siti air itu?
drink-Q Siti water that
'Did Siti drink that water?'

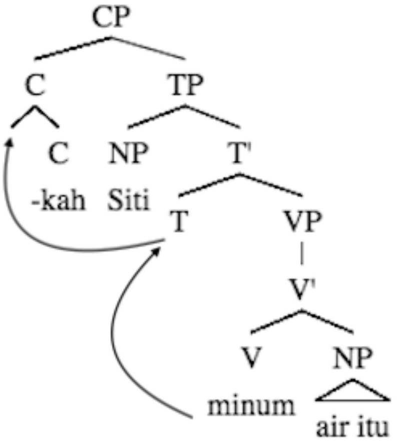


Figure 1. Analysis of Verb Movement.

(20) a. Il embrasse souvent Marie.
he kisses often Mary
'He often kisses Mary.'
(Freidin, 2012, p.156)

b. Embrasse-t-il souvent Marie?
kisses he often Mary
'Does he often kiss Mary?'
(Freidin, 2012, p.156)

Modal movement in Malay polar questions

Further support for such an analysis comes from polar questions when a modal is present. The HMC requires that the verb must first move to C near sentence-initial position via T. However, if T is occupied, such an option becomes unavailable, which is borne out in (21c). Fronting the V head *pergi* will result in an ungrammatical sentence, as there is an intervening T head *mesti* between C and V. Instead, only the modal *mesti* which occupies the T head position can move up to combine *with -kah* as it occupies a higher position than *pergi*.

(21) a. Ali mesti pergi ke sekolah.
Ali must go to hospital
'Ali must go to the hospital.'

b. Mesti-kah Ali pergi ke
must-Q Ali go to
hospital?
hospital
'Must Ali go to the hospital?'

c. *Pergi-kah Ali mesti ke
go-Q Ali must to
hospital?
hospital
'Intended: Must Ali go to the
hospital?'

Based on the discussion above, we have shown that Type 1 polar questions are indeed the result of head movement.

4.2 Phrasal movement

The phrasal movement strategy is another way commonly used to support *-kah* in Malay polar questions. Based on the examples in (15), two types of phrasal movement can be identified in Malay polar questions, i.e.:

1) movement of a phrase to a position before *-kah*
2) movement of a whole sentence to a position before *-kah*

In (15b), the prepositional phrase *ke sekolah* is moved to the beginning of the sentence with the particle *-kah* attached to it to form a polar question. The fact that only the PP *ke sekolah* 'to school' appears before *-kah* shows that *dia* is not moved and remains in its original position.

Although the particle *-kah* in (15c) would appear, at first glance, to be occupying the sentence-final position, we want to argue that, in fact *-kah* remains in its original C position. A plausible explanation for why the particle *-kah* would appear sentence-finally is that the whole sentence, say, TP *dia ke sekolah* is moved and placed before *-kah*, thus giving rise to the illusion that *-kah* moves to the end of the question.

4.3 *ada(kah)* insertion

The last strategy involved in supporting the -*kah* in Malay polar questions is the insertion of *ada* to the sentence-initial position. As illustrated in (16), *ada* is added before *-kah* and followed by the original declarative sentence to give the sentence a question interpretation. Unlike the two strategies mentioned earlier, where the moved element originates from the sentence, *ada* does not originate from the corresponding declarative sentence. Similarly, this can be observed in English *do*-insertion (22). In English, as the main verb cannot move to the T head or C head to form a polar question, when there is no modal or auxiliary verb, the dummy auxiliary *do* is base-generated in T head to support the bound tense morpheme, and then moves to the C head in order to form a polar question in English (Han & Kroch, 2000).

(22) a. *Paint you?

b. Do you paint?

Following this line of thought, we argue that *adakah*-insertion has the same underlying mechanism as *do*-insertion. We propose that *ada* functions similarly as *do* and is base-generated to support *-kah*. However, it is unclear at this stage where *ada* is base-generated. It can potentially be base-generated in T head or C head. However, it seems possible for *ada* to base-generate in the C head.

4.4 How the different strategies can fit into the proposed analysis

Thus far, we have shown that the distribution of *-kah* in Malay polar questions is not arbitrary and have identified the three strategies to support *-kah* in Malay polar questions. We will further show how our proposed analysis is capable of capturing all three strategies. As our proposed analysis is based on the assumption that *-kah* is a bound morpheme that must attach to a host element such as a word or a phrase, it is, therefore, required for something to appear before it. Malay uses any of the three strategies to achieve this. This is evident in the observations made earlier, where the fronted element or *ada* always appears to be adjacent to *-kah*. It should be noted that, however, it is not possible to apply more than one strategies simultaneously.

As *-kah* is known as the question particle, we would assume *-kah* to be occupying the C head with the Q feature. As a bound morpheme, *-kah* is required to attach to a host element in a way similar to the possessive *'s* in English. Therefore, the position before *-kah* must be filled. By examining the deep structure, we can tell the host element must either be a phrase in the Spec of CP or a head that is moved to C or base-generated at C.

In the case of head movement strategy, i.e. Type 1, we argue that the head (be it the T head or V head) will fill the C head position. If T is occupied by a modal, T will move to C instead of V since T has a higher position than V. As C is already occupied by *-kah*, we suggest that the V or T will form a complex C head with *-kah*. Such a movement serves to support *-kah*.

Similarly, we also argue that the phrasal movement in Malay polar questions, i.e. Type 2a and 2b, is motivated by the same reason. We propose that the element that has undergone fronting to be occupying the Spec of CP, as the fronted element is either a phrase or sentence, which is generally assumed to be moved to a phrasal position (Harizanov, 2019).

While the base-generated *ada* strategy is not the result of any syntactic movement, this strategy fits well into the analysis proposed here. In the case where neither head movement nor phrasal movement is involved, it is possible to base-generate *ada* in the C head to support *-kah*. This is not at all unusual as comparable mechanism is also found in English *do*-insertion, where the dummy auxiliary is base-generated at T. See examples in (22). In Malay, *ada* will be base-generated to support the particle *-kah*.

4.5 More on the semantic properties of Malay polar questions

Generally speaking, polar questions formed with head movement have a neutral reading, while polar questions with phrasal movement would receive a focus interpretation. Let us compare (13b) and (15b). As illustrated in (13b), the question remains neutral when only the verb is moved. In contrast, the fronted phrase *ke sekolah* is focus marked in (15b). Note that Malay polar questions that receive focus reading is semantically equivalent to English cleft construction *it is…that…?*. In the same way, the polar question formed in (15c) also receives focus reading as it is also derived by phrasal movement. It is possible to analyze such a movement as phrasal movement if we argue that the entire sentence is moved to a position before *-kah*. It should be noted, however, the constituent that receives focus reading has to be the phrase that is placed closest to *-kah*. As shown in (15c), the phrase that is focus-marked is *ke sekolah*.

Similar to polar questions derived by the head movement strategy, Malay polar questions formed with *adakah* insertion are neutral. Adding *adakah* will only give the sentence a question interpretation, but not a focus reading.

5 Conclusion

In this paper, we have proposed a unified analysis of *-kah* in Malay polar questions. We have identified and examined the different strategies used in Malay polar questions, i.e. head movement, phrasal movement and *ada(kah)* insertion. It has been shown that all the three strategies point to the same requirement, i.e. to support the bound question particle *-kah* at the C head. That is to say something needs to be attached to *-kah* since the bound morpheme cannot stand alone. In the case of head movement and phrasal movement, we have proposed the different landing sites of the fronted element. We have also provided some cross-linguistics evidence to support our claim that *ada* is base-generated to support *-kah* similar to *do*-insertion in English.

References

Fortin, C. (2009). On the left periphery in Indonesian. *Proceedings of the Sixteenth Meeting of the Austronesian Formal Linguistics Association*, 29-43. https://ir.lib.uwo.ca/cgi /viewcontent.cgi?article=1020& context=afla

Han, C. & Kroch, A. (2000). The rise of do-support in English: implication for clause structure. *North East Linguistics Society, 30*(1).

Harizanov, B. (2019). Head movement to specifier positions. Glossa: A Journal of General Linguistics, 4(1), 1-36. https://doi.org/10.5334/gjgl.871

Freidin, R. (2012). *Syntax: Basic concepts and applications*. Cambridge: Cambridge University Press.

Karim, N.K., Onn, F.M., Musa, H.H., & Mahmood, A.H. (2008). *Tatabahasa dewan edisi ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Travis, L. (1984). Parameters and effects of word order variation. (PhD thesis). Massachusetts Institute of Technology.

Abbreviations

ACT	active voice marker
COMP	complementizer
FOC	focus
Q	question particle/marker
TEMP	temporal marker

A distinctive collexeme analysis of near-synonym constructions "*ying-dang/ying-gai* + verb"

Zhuo Zhang
Department of Linguistics
and Translation
City University of
Hong Kong
83 Tat Chee Avenue,
Kowloon, Hong Kong, China
jessyzh@pku.edu.cn

Meichun Liu
Department of Linguistics
and Translation,
City University of
Hong Kong,
83 Tat Chee Avenue,
Kowloon, Hong Kong, China
meichliu@cityu.edu.hk

Dingxuan Zhou
School of Mathematics
and Statistics,
University of Sydney,
Sydney NSW 2006,
Australia
dingxuan.zhou@
sydney.edu.au

Abstract

This paper aims to differentiate the two modal auxiliary near-synonyms *ying-dang* 'should; ought to; shall' and ying-gai 'should; ought to' through investigating their distinctive collexemes of the near-synonym constructions of "*ying-dang*/ *ying-gai* + verb". We employed the distinctive collexeme analysis (Gries and Stefanowitsch, 2004) to find the distinctive verbal collexemes and then categorized them into different semantic types based on the notion of frames (Fillmore, 1982). The sample was extracted from a news corpus of 2.95 billion Chinese characters. The results found that *ying-dang* and *ying-gai* display critical differences in attracted collexemes regarding modality types, usage patterns, and semantic types of verbs. Specifically, *ying-dang* prefers to exhibit a robust deontic meaning, whereas *ying-gai* can display both deontic and epistemic senses. *Ying-dang* also tends to take two-character verbs and appears on formal occasions, whereas *ying-gai* likes single-character verbs better and shows up in various informal contexts. In terms of semantic frames, *ying-dang* is inclined to take verbs with purposive efforts, whereas *ying-gai* attracts verbs of self-motion and emotion.

1 Introduction

Modal auxiliary verbs and adverbs are crucial in understanding the intention of a speaker. The Chinese modals pair *ying-dang* 'shall; ought to; should' and *ying-gai* 'should; shall' is considered as should-modals (similar to should) with overlaps in semantics and functions. Despite native speakers themselves may not be able to explain why they choose one word over another, their choices between near-synonym modals are found to display a distinctive trend in distribution (Hilpert and Flach, 2021). The distribution hypothesis that the semantics of a word is reflected by the company it keeps (Firth, 1957; Harris, 1970: 785; Turney and Pantel, 2010) may offer a potential theoretical foundation for this phenomenon. The collocational preferences can be distinguished between near-synonym constructions (Gries and Stefanowitsch, 2004; Hilpert, 2008) as illustrated by usage-based Construction Grammar (Goldberg, 1995; Diessel, 2019). Also, the divergence of collocational profiles associated with different words could offer vital clues to their differentiated semantics (Hilpert and Flach, 2021).

The subtle usage and semantic differences of modals can also be found through the pair-wise comparison of near-synonym constructions. Hilpert and Flach (2021) found that English modals could be differentiated by their collocational profiles, and near-synonymous pairs like may and might or must and have to differ in distribution. Urunbaevna (2022) employed Distinctive Collexeme Analysis (DCA) to analyze the strong lexemes of construction "should + verb" and "have to + verb" in 3000 sentences of each structure from the British National Corpus. The study found that have to is inclined to verbs with dynamic senses, whereas should prefers stative verbs that tend to appear in the written register and exhibits a stronger obligation sense. However, although "modal + verb" is one of the most representative constructions of modals, few studies have been conducted on differentiating the verbal collexemes between the Chinese near-synonym pairs. Thus, this study intends to compare the distinctive verbal collexemes of the Chinese should modals *ying-dang*/*ying-gai* in the construction of "modal + verb."

Methodologically, this study adopted distinctive collexeme analysis (DCA), one of the emerging quantitative corpus-linguistic methods, to compare the distinctive collexemes between the near-synonym constructions (Gries and Stefanowitsch, 2004; Gries, 2012). As a type of collostructional analysis (Stefanowitsch and Gries, 2003), DCA focuses on usage-based and pattern-specific properties through objective and systematic statistical investigation (Stefanowitsch and Flach, 2020). The method can go beyond the raw frequency to find positively and negatively associated collexemes (Flach, 2020). The application of DCA to the Chinese near-synonymous pair *ying-dang* and *ying-gai* is expected to exhibit a promising result and reflect the subtle semantic features and usage patterns associated with each modal.

Organization of this paper. The rest of this paper is organized as follows. Section 2 presents research aims and questions. Section 3 focuses on previous works. Section 4 illustrates the corpus and method applied in this study. Section 5 analyzes the distinctive verbs associated with the construction of "*ying-dang*/*ying-gai* + verb," and then discusses the limitations and. The last section concludes this paper with the differences between *ying-dang* and *ying-gai*.

2 Research aims and questions

This paper intends to find out the collexeme variances of the Chinese two-character *should* modal pair *ying-dang* and *ying-gai* in the representative construction of "modal + verb" and hopes to illustrate how speakers choose between the two modals. The research questions are:

1) What are the distinctive collexemes of the two near-synonym constructions?
2) How to explain the collexeme preferences of *ying-dang* and *ying-gai* in terms of modality, usage patterns, and semantic types?

3 Previous works

This section presents the previous works on the Chinese modals and introduces the recent advances in distinctive collexeme analysis.

3.1 A comparison of *ying-dang* and *ying-gai*

Ying-dang and *ying-gai* are two frequent two-character modals expressing a similar meaning to *should* in Chinese. They are usually categorized as *gai*-modals in Chinese, which can be translated as modals of *should*', which could both express possibility and necessity (Ding, 1961; Zhu, 1982; Xu, 1990; Chen, 2006; Guo, 2011; Wu, 2021), corresponding to epistemic and deontic modality (Peng, 2007; Pan, 2010). Cao (1999) found that 必须 *bi-xu* 'shall,' *ying- dang* 'should or ought to', 可以 *ke-yi* 'may', and 不得 *bu-de* 'shall not' were often used in legal performatives to impose obligations, conferred rights, permission, and prohibition. The study also found that 必须 *bi-xu* and 应当 *ying-dang* are employed to impose illocutionary force of obligations, whereas *ying-gai* seldom appears in the legislation. Zhou (2008) proposed that *ying- gai* is easier to be adapted to different occasions compared with other modals such as *ying-dang* and *bi-xu*. In comparison, the usage of *ying-dang* is always discussed in legislative texts (Zhao, 2009). Li et al. (2016) studied three pairs of modals with deontic meaning in Hong Kong legislation, including 1) obligation (similar to *shall/must*) 须 *xu*/ 必须 *bi-xu*; 2) no-obligation (similar to *needn't*) *wu-xu*/*bu-bi*; 3) obligation (similar to *should*) *ying-gai*/*ying-dang*. The study found that when lexemes like *ying* and *xu* are encoded in the linguistic context, the semantics must fit in to match the occurrence of *ying* and *xu*, which explains their similarities in meaning. The authors also mentioned that such classification ignored the subtle inner-group differences between *ying-gai* and *ying-dang*. Yao (2017) systematically compared the differences between *ying-dang* and *ying-gai*. The author discovered that 1) *ying-gai* is more flexible in expressing both epistemic and deontic modality while *ying- dang* tends to express deontic modality; 2) when expressing deontic modality, *ying-dang* expresses deontic necessity whereas *ying-gai* deontic obligation; 3) *ying-gai* is more informal than *ying-dang* and can be used in daily conversation; 4) in terms of the constructions of "*ying-gai*/*ying-dang* + 说 *shuo* 'say'," the former is frequently used in the front of a sentence as a summary with a relatively weaker subjectivity. *Ying-gai* and *ying-dang* also displayed distinctively different attractions to prepositions and adverbs (Yao, 2017).

In summary, various studies touched upon the usage patterns of *should* modals in multiple aspects, but few studies have focused on the distinctive collexemes attracted to a specific construction.

3.2 Recent advances in Distinctive Collexeme Analyses

Distinctive collexeme analysis (Gries and Stefanowitsch, 2004) aims to find the more attractive words of one construction as opposed to the other (Hilpert, 2014). One of the examples in this study is the comparison between the near-synonymous construction "*will* + verb" and "*be going to* + verb" denoting the meaning of future. The underlying theoretical foundation is that a verb is only suitable on the condition that its argument matches the construction under the principle of Semantic Coherence (Goldberg 1995: 50). The study found that *will* prefers relatively non-agentive or low-dynamicity actions (*find*, *receive*, *hold*, *finish*, *reach*), including perception/cognition events (*see*, *know*, *want*, *consider*, *notice*, *need*, *accept*), or states (*depend*, *remain*, *become*) whereas the opposite trend was found for *be going to*. DCA is often used to compare collexemes between two or more near- synonym constructions to draw implications on differentiating the near-synonyms (cf. S + *can* + V vs. S + *be able to* + V, Lojanica, 2021)

However, collostructional analysis is not without disputes. Schmid and Küchenhoff (2013) argued that collostructional analysis, including DCA, inherited potential problems in terms of statistical measures used to calculate collocational strengths (mostly Fisher-Yates exact, abbreviated as FYE). Gries (2015, 2019) defended that 1) many association measures, such as G^2 (the log-likelihood ratio), *chi-square*, Mutual Information (MI), and a logged odds ratio, could be used other than the FYE; 2) the FYE was correlated with the most proposed statistical tests and did not influence the top collexemes which were used to report the results; conflation due to large corpus size was itself a feature that describes the frequency of (co-) occurrence, which is important to usage-based grammar.

In a nutshell, DCA is an efficient method for differentiating near-synonymous constructions but rarely used in studying Chinese near- synonyms, and the technique is still developing with continuous trials and improvement.

4 Method

This section presents the corpus and major procedures of this study.

4.1 The corpus

The corpus adopted in this study is *news2016zh* (version 1.0), an open-access corpus with more than 2.95 billion Chinese characters in the large- scale natural language processing Chinese corpus (Xu, 2019). The data was downloaded on March 3, 2022. It contains 2.5 million news articles collected from 63 thousand media platforms between 2014 and 2016. All sentences that contain the structure "*ying-dang* + verb" (CNX1) and "*ying-gai* + verb" (CNX 2) in the corpus were extracted and utilized as samples in this study.

4.2 Major procedures

The major steps are presented as follows:

1. Build the corpus, segment words and perform the POS tagging.
2. Find and extract all near-synonym constructions and use them as the samples of *ying-dang* and *ying-gai*.
3. Count the frequency of each verb respectively in the two samples.
4. Perform the distinctive collexeme analysis.
5. Analyze the top distinctive collexemes by groups and report the results.

The Stanford CoreNLP Natural Language Processing Toolkit (Manning et al. 2014) was used for word segmentation and POS tagging. The collexemes of near-synonym constructions "ying-dang + verb" (CNX1) and "ying-gai + verb" (CNX 2) were counted via self-established programs in the programming language of Java. The frequency was conducted via existing R codes (Flach, 2021) with minor adjustments. FYE was adopted as the association measure as it generally exhibits good performance and correlates with other statistical measures. Also, our focus was on the report of top collexemes rather than the values of the statistical values (Gries, 2015; 2019). The threshold of the statistic measure was decided via multiple trials based on the number of distinctive collexemes with statistical significance. To exclude the low-frequency verbs, we set the minimum sum of CNX1 and CNX2 for each collexeme (referred to as threshold in DCA) as 100 to be considered for DCA, and collexemes that were not verbs were excluded or analyzed during the manual analysis. The analysis of verbal collexemes was based on the notion of frames (Fillmore, 1982) and the grammatical functions of the collexemes. Specifically, we referred to Mandarin VerbNet (Liu, 2017) and FrameNet (Fillmore and Baker, 2010) for deciding the semantic frames of verbs and reported the attracted frames along with some other superficial features associated with the distinctive collexemes.

5 Result and discussion

This section illustrates the descriptive statistics, distinctive collexemes of the two constructions, together with limitations and future implications.

5.1 Descriptive statistics

The corpus contains 161,602 sentences with *ying-dang*, and 548,303 with *ying-gai*, about 3.39times of *ying-dang*. After POS tagging, 41,693 sentences were found tohave the construction of "*ying-dang* + verb" involving 3,057 individual verbs, among which 1,621 verbs appeared over once and 727 more than five times (See Table 1). The corpus also includes 226,309 sentences of "*ying-gai* + verb" (5.43 times of *ying-dang*), which involved 8,876 unique verbs, with 4707 showing up not less than twice and 1,974 over five times. Also, *ying-dang* prefers to take verbs directly than *ying-dang* because the count of *ying-gai* is 3.39 times of *ying-dang* while the usage of "modal + verb" is 4.43 times of *ying-dang*.

Total	*Ying-dang*	*ying-gai*	*ying-gai/ ying-dang*
sent	161,602	548,303	3.39
CNX	41,693	226,309	5.43
verbs	3,057	8876	2.90

Table 1: Descriptive statistics

5.2 Most distinctive collexemes in CNX 1

After the DCA, 281 out of 3,027 individual verbs met the threshold, 110 verbs are distinctive to *ying-dang* with statistical significance in CNX 1, and 61 verbs are not distinctive to *ying-dang* or *ying-dang*. The top 20 distinctive collexemes associated with CNX1 are presented in Table 2 on the next page, together with observed and expected frequencies of the two CNXs, collocation strength, and significance level of p < 0.0001. *Ying-dang* is more frequently used in formal occasions associated with deontic needs in an obligation or requirement scenario, such as laws, regulations, official announcements, policy, obligations, and requirements, as shown in Example (1)-(6), in which the typical verbs are 遵守 *zun-shou* 'obey', 要求 *yao-qiu* 'require,' 允许 *yun-xu* 'permit,' 承认 *cheng-ren* 'admit,' 负责 *fu-ze* 'account for', and 禁止 *jin-zhi* 'forbid.' *Ying-dang* are inclined to take two-character verbs from various semantic types (frames) with a strong inclination of purposive efforts (See Table 3).

Semantic frames	Typical verbs	E.g.
Purposive efforts	遵循 *zun-xun* 'obey' 履行 *lv-xing* 'perform (duties)'	(1)
Existence	具备 *jv-bei* 'possess' 具有 *jv-you* 'possess'	(2)
Cognition	坚持 *jian-chi* 'insist' 视为 *shi-wei* 'consider (as)'	(3)
Communication	说明 *shuo-ming* 'state' 告知 *gao-zhi* 'tell'	(4)
Transference	支付 *zhi-fu* 'pay'	(5)
Attribute of manner	真实 *zhen-shi* 'real' 谨慎 *jin-shen* 'cautious'	(6)

Table 3: Representative verbs and frames in the construction of "ying-dang + verb" (CNX1)

No	LEX	*pinyin*	Eng	O.CXN1	E.CXN1	O.CXN2	E.CXN2	COLL	SIGNIF
1.	承担	*cheng-dan*	undertake	1485	381.6	968	2071.4	Inf	*****
2.	符合	*fu-he*	conform	1022	186.5	177	1012.5	Inf	*****
3.	认定	*ren-ding*	identify	574	95.4	39	517.6	Inf	*****
4.	提交	*ti-jiao*	submit	492	80.1	23	434.9	Inf	*****
5.	予以	*yu-yi*	give	599	120.1	173	651.9	Inf	*****
6.	遵守	*zun-shou*	obey	578	125.5	229	681.5	277.5758	*****
7.	建立	*jian-li*	Establish	697	179.8	459	976.2	262.8857	*****
8.	具备	*jv-bei*	have	880	321.6	1187	1745.4	189.5545	*****
9.	履行	*lv-xing*	fulfill	326	62.7	77	340.3	185.423	*****
10.	取得	*qu-de*	get	253	45	36	244	161.308	*****
11.	提供	*ti-gong*	supply	346	92.6	249	502.4	124.0307	*****
12.	遵循	*zun-xun*	follow	422	135.2	447	733.8	114.3772	*****
13.	组织	*zu-zhi*	organize	165	29.4	24	159.6	105.0159	*****
14.	采取	*cai-qu*	take	504	195.1	750	1058.9	97.49452	*****
15.	说明	*shuo-ming*	illustrate	168	32.8	43	178.2	93.84047	*****
16.	包括	*bao-kuo*	include	420	153.2	565	831.8	90.92773	*****
17.	载明	*zai-ming*	state	112	17.4	0	94.6	90.55912	*****
18.	真实	*zhen-shi*	real	116	18.5	3	100.5	88.57565	*****
19.	加强	*jia-qiang*	amplify	439	169.3	649	918.7	85.66282	*****
20.	披露	*pi-lou*	reveal	129	23	19	125	82.06816	*****

Note: O.CXN = Observed Construction Frequency; E.CXN = Expected Construction Frequency

Table 2: The most distinctive verbs in the construction of “ying-dang + verb

Example (1) 制定和实施城乡规划**应当遵循**…的原则。

The formulation and implementation of urban and rural planning **shall follow** the principles of …

Example (2) 申请认定小学教师资格，**应当具备**高等院校专科毕业及其以上学历。

To apply for the qualification of primary school teachers, one **should have** a college degree or above from an institution of higher learning.

Example (3) 行政机关依申请公开政府信息**应当坚持**公正、公平、便民、及时的原则

When an administrative organ discloses government information upon application, it**shall adhere to** the principles of impartiality, fairness, convenience, and timeliness.

Example (4) 报告期内发生重大会计差错更正需追溯重述的，公司**应当说明**情况、更正金额、原因及其影响。

Suppose a significant accounting error correction occurs during the reporting period thatneeds to be retrospectively restated, the company**shall explain** the situation, correct the payment amount, and indicate the reason and its impact.

Example (5) 劳动合同法第八十五条进一步明确了法律责任，劳动报酬低于当地最低工资标准的，**应当支付**其差额部分;

Article 85 of the Labor Contract Law further clarifies the legal responsibility. If the labor remuneration is lower than the local minimum wage, the difference part **shall be** paid;

Example (6) 报考人员提交的报考申请材料**应当真实**、准确

The application materials submitted by the candidates **should be true** and accurate.

Ying-dang also prefers two-character light verbs (see Table 4 on the next page), which add little semantic meaning to the sentences other than bring or amplify the following verbs.

Apart from light verbs, the construction of “verb of attaining/achieving/adopting + goal” is also frequently occurred with *ying-dang*, in which the meaning is largely reliant on their objects (See Table 5 on the next page).

Example (7) 该通告针对蓬江、江海、新会三区有效，要求其范围内销售燃气具的单位**应当取得**营业执照并注明相应的经营范围。

Example (7) This circular is valid for the three districts of Pengjiang, Jianghai, and Xinhui. and it requires that units selling gas appliances in these areas should obtain a business license and indicate the corresponding business scope.

Semantics	Light verb	Literal meaning	E.g.
To bring the following verbs	予以 *yu-yi* 给予 *ji-yu*	give	予以赔偿 'to be compensated'
Existence	作出 *zuo-chu* 进行 *jin-xing* 实行 *shi-xing* 加以 jia-yi	do/ perform	作出无罪判决 'to be acquittal'
Cognition	引起 *yin-qi*	cause	引起警惕 'to arouse vigilance'
To amplify the following verbs	加强 *ji-qiang* 加大 *jia-da*	amplify/ strengthen	加强中外交流 'to strengthen Sino-foreign exchanges'

Table 3: The distinctive light verbs in "*ying-dang* + verb

Semantic frames	Typical verbs	E.g.
attaining	取得 *qu-de* 'obtain' 获得 *huo-de* 'achieve'	(7)
achieving	达到 *dao-da* 'reach' 发挥 *fa-hui* 'play a role' 满足 *man-zu* 'satisfy'	(8)
adopting	采用 *cai-yong* 'use' 采取 *cai-qu* 'take'	(9)
Communication	说明 *shuo-ming* 'state' 告知 *gao-zhi* 'tell'	(4)

Table 4: Typical verbs belong to verbs of attaining/achieving/adopting

Example (8) 普通话水平**应当达到**国家语言文字工作委员会颁布的普通话水平测试等级标准二级乙等以上**标准**。

Example (8) The proficiency of Putonghua **should meet the standard of** Level 2 or above of the Putonghua Proficiency Test issued by the State Language and Character Work Committee.

Example (9) **应当采用**反担保等必要**措施**防范风;

Example (9) Necessary **measures,** such as counter-guarantee, **should be adopted** to prevent risks

5.3 Most distinctive collexemes in CNX 2

Compared with ying-dang, ying-gai tends to take a wider semantic range of verbs with no bias on single-characters verbs. The top 20 distinctive collexemes are presented in Table 7 on the next page. Compared with ying-dang, verbs of motion transference, perception, and emotion are more likely to co-occur with ying-gai. Some of the representative verbs are summarized in Example (10)-(15) in Table 6.

Semantic frames	Typical verbs	E.g.
Self-Motion	去 *qu* 'go' 来 *lai* 'come'	(10)
Communication	说 *shuo* 'say'	(11)
Cognition	懂 *dong* 'understand' 想 *xiang* 'think' 知道 *zhi-dao* 'know' 明白 *ming-bai* 'understand'	(12)
Existence	有 *you* 'have/ possess'	(13)
Transference	买 *mai* 'buy' 花 *hua* 'spen 学 *xue* 'learn' 换 *huan*'exchange'	(14)
Emotion	高兴 *gao-xing* 'happy'	(15)

Table 5: Representative verbs and frames in the construction of "ying-gai + verb" (CNX2)

Example (10) 你若问我中国哪里最**应该去**，我会告诉你新疆!

If you ask me which is the best place people **should go** to in China, I will tell you Xinjiang!

Example (11) 这句话似乎**应该说**成我们现在都是老兵了。(Literal meaning: should say)

It seems this sentence should be put in this way: we are all veterans now.

Example (12) 他爱我，就应该懂我、满足我。

He loves me, so he should understand me and satisfy my needs.

Example (13) 他是第一个听我哭的男人，他教我男人就**应该有**强壮的臂弯。

He is the first man to hear me cry, and he teaches me that men **should have** a strong arm.

Example (14) 在跑步一开始，**应该买**些最好的鞋子和衣服吗?

Should we **buy** the best shoes and clothes as soon as we begin to run?

Example (15) 你见到我**应该高兴**呀。

You **should be happy** when seeing me.

It was also observed when the typical perception verb 看 kan 'see/look at' is used after ying-gai, it tends to express abstract cognitive activities, such as focusing on, checking, valuing, reading something or the indicated meaning of seeing a doctor. The respective examples (extracted sentence clips) can be found in Example (16)-(20).

Example (16) **应该看的是**它的潜藏价值

'**should focus on** its potential value'

Example (17) 应该**看两证**

'**should check** two certificates

Example (18) **应该看得**比生命还重要

'should value (it) over than life'

Example (19) 应该看过精益创业 'should read *The Lean Startup'*

Example (20) 应**该看**什么科?

Word-for-word: Should – see – what – section? 'Which section should (a potential patient) go?'

Among the top 50 verbs distinctive to ying-gai, more a half are single-character verbs, whereas those to ying-dang are all two-character verbs. Ying-gai can be considered more colloquial than ying-dang, evidenced by the frequent uses of single-character verbs and informal expressions like trifling in daily chat. For example, informal verbs like 叫 jiao 'call … as' and 算 suan 'is barely considered as' were found distinctive to ying-gai as shown in Example (20)-(21).

Example (21) 职称药师考试其实**应该叫**药学职称考试

The Professional Pharmacist Qualification Exam **should** actually **be called** the Pharmacy Job Title Exam.

No	LEX	*pinyin*	Eng	O.CXN1	E.CXN1	O.CXN2	E.CXN2	COLL	SIGNIF
1.	说	*shuo*	say	144	1122	7068	6090	Inf	*****
2.	是	*shi*	be	2644	9185.6	56401	49859.4	Inf	*****
3.	会	*hui*	can/will	131	1029.1	6484	5585.9	309.0393	*****
4.	做	*zuo*	do	117	746.6	4682	4052.4	203.3121	*****
5.	有	*you*	have	882	1822.3	10832	9891.7	156.5101	*****
6.	要	*yao*	will	124	564.6	3505	3064.4	125.4005	*****
7.	去	*qu*	go	96	438.2	2721	2378.8	97.55963	*****
8.	算是	*suan-shi*	be considered	31	244.4	1540	1326.6	73.33712	*****
9.	可以	*ke-yi*	can	67	318.9	1983	1731.1	73.2193	*****
10.	能	*neng*	can	59	287.8	1791	1562.2	67.22511	*****
11.	怎么办	*zen-me-ban*	how to do	25	198	1248	1075	59.65883	*****
12.	让	*rang*	let	171	433.1	2613	2350.9	52.67851	*****
13.	没有	*meiyou*	no	22	159.9	1006	868.1	46.60287	*****
14.	没	*mei*	no	5	107.3	685	582.7	42.28269	*****
15.	给	*gei*	give	110	291.2	1762	1580.8	37.9745	*****
16.	叫	*jiao*	call	13	105.3	664	571.7	32.31799	*****
17.	知道	*zhi-dao*	know	306	511.5	2982	2776.5	25.73305	*****
18.	算	*suan*	count	15	92.7	581	503.3	25.32514	*****
19.	明白	*ming-bai*	clear	45	146.5	897	795.5	24.87238	*****
20.	买	*mai*	Buy	2	57.6	368	312.4	23.81272	*****

Table 6: The most distinctive verbs in the construction of "*ying-gai* + verb"

Example (22) 如果孩子写的是 make, 那**应该算**对还是错呢?

If children write ‘make,’ **should it be considered as** right or wrong?

Ying-gai also prefers highly transitive constructions over *ying-dang*, evidenced by 让 *rang* /给 *gei* construction ‘enable …’ as shown in Example (23) and frequent usage of caused-motion/position verb 放 *fang* ‘put’ as presentedin Example (24).

Example (23) 我们**应该让它恢复**到自然的规律当中去。

We **should enable it to restore** the law of nature.

Example (24) 主要精力**应该放**在教育教学管理上。

The primary energy **should be put** intoeducation and teaching management.

Apart from these examples, *Ying-gai* also tends to appear in front of modal auxiliary verbs. Among the top 10 distinctive collexemes, fourare modals, including 会 *hui* ‘can/will’, 要 *yao* ‘want’, 可以 *keyi* ‘can’ and 能 *neng* ‘can’.Compared with those of*ying-dang*, the distinctive collexemes of*ying-gai* could already display epistemic probabilities as shown inExample (23) and (24).

Example (23) 志玲姐姐的导航声音嗲嗲的，**男司机应该会喜欢**。

Female actor Zhiling's navigation voice is squeaking, so **male drivers should like it.**

Example (24) 但是范冰冰私生子的谣言**应该可以不攻自破了**。

However, the rumor of Bingbing Fan's illegitimate child **should be self-defeating**.

Similarly, the distinctive collexemes of*ying- gai* also include interrogative phrase 怎么办 *zen-me-ban* ‘how to deal with it’, negation adverb negation 没 / 没有 *mei*/*mei-you* ‘not/haven’t’though these words are not considered as verbs.

5.4 Limitations and future implications

In terms of the research scope, this study only focuses on the distinctive collexemes associated with the near-synonym constructions “*ying-dang*/*ying-gai* + verb”. To fully address their differences, a future study may explore more constructions, such as “modal + adv,” “adv + modal,” and “modal + prepositions”. In the corpus, we also found that “Degree markers + modal” is much less frequently collocated with *ying-dang* rather than *ying-gai* (Yao, 2017). As shown in Example (10), degree marker 最 *zui* ‘the most’ is used before *ying-gai*. Flach (2020) employed Co-Varying Collexeme Analysis, also a type of collostructional analysis, to study gradient idiomaticity in MOD + ADV collocations, and revealed that collocational behaviors of modal auxiliaries could serve as a cue for measuring the scope of adverbial modification.

Besides, this study mainly relies on POS tagging to extract the target constructions, which was hardly 100% accurate. However, some improper tags may still be meaningful as the same POS model was applied to the samples. It is suggested to keep the original data as much as possible while pointing out the potential improper tags, especially when such tagging could also reveal differences in usage patterns. The point is that loyalty to the actual data is essential, and ‘wrong’ or unexpected results should not be ignored. Ignoring the ‘wrong’ or unexpected tags may miss critical aspects in the findings. In the case of this study, modal auxiliary verbs, such as 会 *hui* ‘can/will’, 要 yao ‘want’, 可以 *ke-yi* ‘can’, interrogative phrase 怎么办 ‘how to deal with it’, and negation adverb negation 没/没有 ‘haven’t’, were found to be distinctive to *ying-gai*. To the authors’ knowledge, such preferences to different modals were first proposed to differentiate *ying-dang* and *ying-gai*.

Also, when analyzing the collexemes, it is suggested to draw some implications on the existing theories like semantic domains (Biber, 1999:365-371) as illustrated by Deshors (2017) and the notion of frames (Fillmore, 1982) evidenced by Wiliński (2019) to provide a relatively objective and linguistically-wise interpretations on semantic types. In fact, the uses of semantic types and frames are not new but seldomly cited in terms of specific theories (Gries and Stefanowitsch, 2004). In the case of this project, we were largely reliant on the frame- based interpretations of verbs.

Last, methodologically, DCA is not just limited to the constructions associated with near- synonyms and can potentially be applied to more construction-based studies. For instance, Newman (2021a) applied this approach to study the differences between singular *child* and plural form *children* by comparing three typical constructions, such as (a) adjective + child/children, (b) child’s/children’s +

noun, and (c) child/children + present participle of a verb. In the same year, Newman (2021b) also employed this approach to study the collexemes distinctive to singular and plural forms of animal nouns *dog* and *cat*. The studies on singular and plural forms are hardly considered as near-synonyms; however, the purpose of this approach is to find the more attractive or repelled collexemes distinctive to the typical constructions of the studied words so as to reveal their usage variances and semantic differences

6 Conclusion

In sum, this paper presented the collexemes respectively distinctive to *ying-dang* and *ying-gai* in the construction of "modal + verb". In terms of modality, the study finds *ying-dang* prefers to take collexemes with a stronger obligation sense, whereas *ying-dang* prefers common verbs with no obvious indication of deontic senses and is able to be used together with other modal verbs to express epistemic meaning. As for usage patterns, *ying-dang* is frequently appeared in a formal context where obligations or requirements are involved and likes to take two-character verbs, whereas *ying-gai* attracts single-character verbs, and appears in various occasions including but not limited to daily conversation, gossips, and forums. Regarding semantic preferences, *ying-dang* prefers verbs associated with purposive efforts whereas *ying-gai* enjoys collocating with verbs of self-motion and emotion. *Ying-gai* is also more likely to take interrogative phrase and negation adverb negation. Methodologically, this study offers some practical suggestions for applying DCA to study Chinese modal auxiliary verbs.

Acknowledgement

The research is supported by the CityU of HK grant for Project #7020010 - Developing a Linguistically-wise and Computationally Feasible Digital System of Semantic Representation for NLP Applications.

References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Longman, Harlow, UK.

Deborah Cao. 1999. 'Ought to' as a Chinese Legal Performative. *International Journal for the Semiotics of Law*, vol. 12, no. 2, , pp. 151–167.

Sandra C Deshors. 2017. Zooming in on Verbs in the Progressive: A Collostructional and Correspondence Analysis Approach. *Journal of English Linguistics*, vol. 45, no. 3, pp. 260–290.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge University Press, Cambridge, UK.

Charles J. Fillmore and Collin F. Baker. 2010. A Frames Approach to Semantic Analysis. In B. Heine and H. Narrog (Eds.) *The Oxford Handbook of Linguistic Analysis. Oxford University Press*. Oxford, UK/New York, New York.

John R. Firth. 1957. A synopsis of linguistic theory 1930–55. Reprinted in Frank. R. Palmer (Ed.), (1968). *Selected Papers of J.R.* Firth 1952–1959. Longman, London, UK.

Susanne Flach. 2021 Collostructions: An R implementation for the family of collostructional methods. Package version v.0.2.0, URL: https://sfla.ch/collostructions/. ED: 3 March. 2022.

Susanne Flach. 2021. Beyond Modal Idioms and Modal Harmony: a Corpus-Based Analysis of Gradient Idiomaticity in Mod Adv Collocations. *English Language and Linguistics*, vol. 25, no. 4, pp. 743–765., doi:10.1017/S1360674320000301.

Adele E. Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago University Press, Chicago, USA

Stefan Th. Gries 2012. Corpus linguistics: Quantitative methods. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, 1380-1385. Oxford: Wiley-Blackwell.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics*, 9.1: 97-129.

Martin Hilpert. 2008. *Germanic Future Constructions. A Usage-based Approach to Language Change*. John Benjamins, Amsterdam, the Netherlands.

Martin Hilpert. 2014. Collostructional analysis: Measuring associations between constructions abd lexical elements. In D. Glynn and J. A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy* (Human Cognitive Processing, vol. 43) (pp.391– 404). John Benjamins, Amsterdam, the Netherlands.

Martin Hilpert and Susanne Flach. 2021. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities*, 36(2): 307–321.

Jian Li, Le Cheng, and Winnie Cheng. 2016. Deontic meaning making in legislative discourse. *Semiotica*, 209 (2016): 323-340.

Meichun Liu. 2017. A frame-based morpho-constructional approach to verbal semantics. In Chunyu Kit and Meichun Liu (Eds.) *Empirical and Corpus Linguistic Frontiers*, China Social Sciences Press, BJ.

Tošić T. Lojanica. 2021. Exploring present ability: A collostructional approach. *Nasledje Kragujevac* 18.48:105-115.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Association for Computational Linguistics, Baltimore, Maryland.

John Newman. 2021a. Child and children in a corpus of American fiction: Contrasting semantic preferences and their experiential motivations. *Cognitive Semantics*, 7.1: 1-30.

John Newman. 2021b. Singular and plural preferences among adjectival collocates of CAT and DOG. *LaMiCuS,* 5: 12-32.

Anatol Stefanowitsch and Susanne Flach. 2020. Too big to fail but big enough to pay for their mistakes:A collostructional analysis of the patterns [too ADJ to V] and [ADJ enough to V]. In Gloria Corpas and Jean Pierre Colson (eds.), *Computational and corpus-based phraseology*, (pp. 248–272). John Benjamins, Amsterdam, the Netherlands.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8.2: 209-243.

Peter. D. Turney and Patrick Pantel (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37: 141–88.

Bright Xu. 2019 Sep. "NLP Chinese corpus:Largescale Chinese corpus for NLP." Corpus version 1.0, URL: https://doi.org/10.5281/zenodo.3402023. ED: 10 March. 2022.

Jiajia Chen. 2006. Multi-perspective research on "*ying-gai*". Hunan Normal University. Ph.D. thesis.

Shengshu Ding. 1961. *Speeches on Modern Chinese Grammar*. Commercial Press. BJ.

Zhaojun Guo. 2011. Two modal types of “gai-type” auxiliary verbs and their selection factors. *Nankai Journal of Linguistics*, (11):132-149.

Lizhen Peng. 2007. *Research on modality of modern Chinese*. China Social Science Press. BJ.

Tiantian Wu. 2021. *A Comparative Study of the Modal Verbs "ying," "gai," "dang," "ying-gai," and "ying."* Shanghai International Studies University, MA thesis.

Heping Xu. 1990. A Preliminary Study on Chinese Modal Verb Semantics and Syntax. World Society for Chinese Language Teaching, *Selected Papers of the 3rd International Conference on Chinese Language Teaching*, Beijing Language Institute Press, BJ.

Wenbiao Yao. 2017. A Comparative Study of "*ying-gai*" and "*ying-dang*". Central China Normal University, MA thesis.

Yancai Pan. 2010. Analysis of the meaning of "*ying-gai*". *Chinese Knowledge*, (04):59-61.

Wei Zhao. 2009. Language Modal Expressions and Its Standardization. *Rhetorical Learning*, (02):30-36.

Youbin Zhou. 2008. Standards of Parts of Speech in Chinese and Determination of Auxiliary Verbs. *Journal of Huaibei Vocational and Technical College*, (06):50-52.

Dexi Zhu. 1982. *Lecture Notes on Grammar*. Commercial Press. BJ.

A Critical Discourse Analysis of Philippine Political Advertising

Heherson P. Valdez
Isabela State University
Cauayan City, Isabela, Philippines
hehersonpasionvaldez@gmail.com

Jayson Mark G. Colcol
Isabela State University
Cauayan City, Isabela, Philippines
jaysonmark14@yahoo.com

Abstract

The growing number of studies in Critical Discourse Analysis, particularly advertisements, highlighted how a particular discourse or group of discourses change or alternate society's traditions, cultures, values, beliefs, norms, and ideologies. Incorporating the framework of Fairclough (2001) and by adopting the transdisciplinary approach of Critical Discourse Analysis, wherein the theories in sociology were integrated into the analysis of texts, this study examines the forty (40) political campaign advertisements from Facebook and YouTube. It analyses the interrelationship of text and context, identifying their underlying sociological and political ideologies. More specifically, this paper examines the Philippine political campaign ads focusing on rhetorical tools, capitals, and representations. There are seven (7) rhetorical tools identified in this study: the use of taglines, testimonials, catchy lines and slogans, the use of humor, figures of speech such as simile and hyperbole, repetition, and ad hominem. The use of two types of capital, social and cultural capital, was also observed in the ads.

On the other hand, three (3) representations of politicians' positionality are identified: politicians as our only hope, politicians as a savior, and politicians as people who grew up in poverty. The samples that had been analyzed show hegemonic features of Philippine political advertising. This study proposes that the creation of political campaign advertisements in the Philippines, as supported by the theory of Bourdieu (1986) and Gramsci (1978), needs to reassess and evaluated.

Introduction

Political discourse analysis (PDA) is concerned with the study of political discourse. PDA is a critical enterprise as well as an arena for political discourse. This would indicate that critical-political discourse analysis, in the vein of contemporary approaches in critical discourse analysis (CDA), focuses on the reproduction of political power, power abuse, or domination through political discourse, including various types of discursive dominance. Such an examination focuses on the discursive circumstances and repercussions of social and political inequality that arises as a result of such dominions (Fairclough, 1995; Van Djik, 1993).

CDA is an interdisciplinary form of analysis as what Fairclough (1995) shall prefer to call a transdisciplinary form of research. This term entails is that the 'dialogues' between disciplines, theories, and frameworks that take place in doing analysis and research are the source of theoretical and methodological developments within the particular disciplines, theories, and frameworks in dialogue, including CDA itself. In this study, CDA was used as the main method of analysis. It also integrates Bourdieu's (1986) Theory of Capital and Gramsci's (1978) Concept of Hegemony. Discourse analysis includes more than just text analysis. Fairclough (2001) recognized three parts of discourse: text interaction, social context, and the corresponding distinction of Critical Discourse Analysis: description of the text, interpretation of the text-interaction relationship, and explanation of the interaction-social context link. The description refers to the stage concerned mainly with the formal properties of the text. Interpretation is the stage concerned with the relationship between the interactions with the text. The latter is seen as the end product of the process of text production and as a resource in the process of text interpretation and explanation, the stage that deals with the relationship between social context and interaction with the social determination of the processes of production and interpretation, and their social effects.

This study integrated two theories in Sociology which are Bourdieu's (1986) Theory of Capital and Gramsci's (1978) Concept of Hegemony. Bourdieu’s (1986) theory of capital has three basic forms, depending on the field it operates and the cost of the more or less expensive transformations required for its efficacy.

These three forms are economic, social, and cultural capital. Another theoretical approach used in this study is concept of hegemony wherein it refers to the ideological superiority of the dominant class's cultural norms, values, and ideas over the subordinated in Gramsci's matrix.

This study analyzed the political discourses here in the Philippines through an in-depth analysis of the senatorial and presidential political ads examining the social formation and representation in these political ads through the integration of Fairclough's (2001) Critical Discourse Analysis framework supported by Bourdieu's (1986) Theory of Capital and Gramsci's (1978) Concept of Hegemony.

Methodology

The researcher gathered the corpus of the research from reliable pages and channels found on Facebook and YouTube. Only selected senatorial and presidential campaign ads from the 2007 to 2019 elections were used in this study. The nature of the research design is purely qualitative as the Philippine political text is analyzed based on Fairclough's CDA framework. There's a large number of televised political ads uploaded on Facebook and YouTube, that's why the researcher chose one political ad for each candidate. The researcher limits the data to thirty (30) senatorial political campaign ads and ten (10) presidential campaign ads.

Verbal texts from the televised ads are the main focus of this study. Each advertisement was transcribed to written data. Fairclough’s (2001) Critical Discourse Analysis (CDA) framework was used to examine the rhetoric elements and identify the capitals and representations present in each advertisement. CDA requires three levels of analysis: 1) Description, 2) Interpretation, and 3) Explanation. Since CDA is an interdisciplinary form of analysis as what Fairclough (1995) shall prefer to call it as a transdisciplinary form, this study will integrate Bourdieu’s (1986) Theory of Capitals and Gramsci’s (1978) Concept of Hegemony. The first level, *description*, refers to the evaluation of rhetorical tools/elements in the ads. The second level, *interpretation*, deals with the analysis of capitals in the ads based on Bourdieu’s (1986) Forms of Capitals. On the third level, *explanation*, the study will utilize Gramsci’s (1978) Concept of Hegemony to inspect the ideologies and describe the representation of politicians’ positionality and social classes’ positionin the ads.

Analysis

Rhetorical Tools in the Ads

An in-depth examination of the rhetorical tools of the senatorial and presidential political campaign ads is crucial in applying the first stage of Fairclough’s CDA (2001) model, *description.* This stage describes the creation and manipulation of rhetorical elements or tools as one of the strategies of persuasions of the politicians in the select political campaign ads.

i. The use of Tagline

One of the unique characteristics of Philippine political advertising, whether local or national election, is the use of tagline. The tagline is a concise phrase that is easy to remember and usually appears at the end of the ad, aiming to complete the explanation of the creative idea of the ad (Moriarty et al. 2014 as cited in Ilhamsyah & Herlina 2019).

Vilma Santos Recto: Ngayong halalan, bumoto potayo ng “correcto”.
(Advertisement #2: Ralph Recto 2007)

The excerpt above shows the tagline of senatorial candidate Ralph Recto. The candidate created his tagline by forming a new word, *“correcto,”* using the lexeme *“correct”* and his last name, *“Recto.” His wife’s statement* above aims to encourage the audience to vote correctly orwisely.

This finding supported what Rahmawati and Tajuddin (2020) found in their study of politicians' billboards in Indonesia that included taglines and messages intended for voters. They claimed that a billboard serves not just to advertise something but also to communicate the creator's identity and ideas. The use of taglines in the politicians’ ads does not promote their names, but it also contains a message, platform, and advocacy.

ii. Testimonial

On traditional screen advertising such as commercials or theatrical ads, they used the testimonial technique to convey a promotional message, either by employing an anonymous model or actor for presenting a product or for showing a product in use or by employing a celebrity, to the effect of integrating personal appeal into an otherwise short, conventional advertising format (Vonderau & Zimmermann, 2021).

Participant 1: Namasukan ako bilang OFW
Participant 2: Nagkaroon ako ng isang abusadongamo
Participant 3: Naging impyerno ang buhay ko
Participant 4: Akala ko diko na muling makita angaking pamilya.
Participant 2: Pero isang araw, sinagip kami ngisang nagmamalasakit na ating kababayan.
Participant 2: Salamat po, salamat sa pangalawang pagkakataon.
Participant 3: Niligtas niya kami mula sa kalupitan na aming pinagdadaanan.
Participant 1: Binigyan niya kami ng pangalawangpagkakataon sa buhay.
(Advertisement #28: Jinggoy Estrada 2019)

The excerpts above show the political campaign ads of senatorial aspirants Jinggoy Estrada and Sonny Angara. Testimonies from the Overseas Filipino Workers (OFWs) and celebrities were used to highlight the name of the candidates. The ad of senatorial candidate Jinggoy Estrada focused on the testimony of the OFWs who narrated their worst experiences while working abroad. After narrating their testimonies, these participants in the ads introduced the candidate as the hero and the main subject of their stories. Campaigns succeed in part by appealing to emotions, and emotional appeals might encourage democratically acceptable behavior (Brader, 2005). The use of testimonies also evokes emotions that make the ads more appealing.

iii. Catchy Lines/Slogans

Aside from using tag lines as a rhetorical tool of persuasion, the use of catchy lines or slogans was also observed in the political campaign advertisements of the senatorial and presidential candidates. Safire (1978), as cited in Newsome (2002), defined a slogan as "a rallying cry; catchphrase; a concise message that crystallizes an idea, defines an issue, the best of which delight,exhort, and inspire."

Music Background Lyrics: Nakaligo ka na ba sa dagat ng basura? Nagpasko ka na ba gitna ng kalsada? Iyan ang tanong namin, tunay ka bang isa sa amin? Nalaman mo na bang mapapag-aral ka niya? Tutulungan tayo para magkatrabaho at ang kanyang plano’y magkabahay tayo! Si Villar ang tunay na mahirap, si Villar ang tunay na may malasakit, si Villar ang may kakayahan at gumawang sariling pangalan! Si Manny Villar ang magtatapos ng ating kahirapan!
(Advertisement #38: Manny Villar 2010)

The excerpt above shows the political campaign ad of presidential candidate Manny Villar. The catchy lines or slogans found in the ad of presidential candidate Manny Villar were found in the first, second, and third lines: “*Nakaligo ka na ba sa dagat ng basura? Nagpasko ka na ba gitna ng kalsada? Iyan ang tanong namin, tunay ka bang isa sa amin? (“Have you ever tried to take a bath in a sea of garbage? Have you ever experienced celebrating Christmas in the middle of the road? This is our question, are you really one of us?).* These lines caught the public’s attention through questions or interrogative statements in the song’s lyrics where the characters who sang this song are “*mga batang lansangan”* (*street* *children*). The ad described their way of living, and it questions the audiences' minds if they can relate to them or not. Barry (1998), as cited in Newsome (2002), argued that the effectiveness of a political slogan is determined by public acceptance. He also said that slogans could cross party lines and be adopted by the entire country. These catchy lines and slogans capture the reality in our society. Still, it could be harmful if the viewers are encouraged to vote for the candidate because of the emotions or dramatic tone residing in the slogans or catchy lines.

iv. The use of Humor

One of the popular beliefs about Filipinos is that they are known for their ability to smile and laugh despite difficulties and has been regarded as a coping mechanism for Filipinos. Humor could have been a coping technique for the Filipinos in the face of adversity (Balmores, 2018).

Participant 1: Walang tatalo sa Tito ko!
Palinawan ng mata? Malayo pa lang kitang kitana, pati kinabukasan mo!
Joey De Leon: Tigilan niyo yang mga tito tito niyo!
Wala yang mga tito niyo dito sa titong ito!
Participant 1: Eh yan din yung Titong sinasabi koeh!
Tito Sotto III: Natural sating mga Pilipino ang masaya sa kabila ng hirap, ang maganda samahannatin ng pagkakaisa ang saya.
(Advertisement #1: Tito Sotto 2007)

As illustrated above, both ads used humor as a rhetorical tool. Senatorial candidate Tito Sotto used this persuasion strategy to instill the value of happiness and unity in the face of adversity. The ad was presented in a dialogue form where the participants boast their *"Titos"* (*Tito* is a Filipino term for uncle) in a witty way; then the candidate appeared in the last scene saying, *"Natural sating mga Pilipino ang masaya sa kabila ng hirap, ang maganda samahan natin ng pagkakaisa ang saya" (It is natural for us as Filipinos to be happy amid adversity, but it is good when combining happinesswith unity).*

According to Young (2017), "political humor" refers to any hilarious text that deals with political concerns, persons, events, procedures, or institutions. Freud (1960), as cited in Harris (2009), argued that jokes could be divided into two basic categories based on the teller's intent. These are referred to as "tendentious" jokes or jokes with a purpose and "innocent" jokes, or jokes with no purpose other than making another party laugh. The finding above revealed that joke was humorously employed in the ad of senatorial candidate Tito Sotto. This is described as a joke with purpose because the political actor's nickname was utilized. At the end of the ad, he delivered his message about happiness and unity amid adversity.

v. Figures of Speech

Figures of speech are used in rhetoric to help orators and politicians shape their discourse and create effective meaning (Alruzzia & Yunusb, 2019). Philippine political advertising also contains figures of speech that help the political actors to input their propaganda and ideologies logically.

The following excerpts show how political actors use the figures of speech as a rhetorical tool forpersuasion.

a. Simile

A simile is a figure of speech in which two essentially distinct objects or concepts are compared using the words "like" or "as." Similes and metaphors are often used together because the simile conveys an explicit comparison while the metaphor expresses an implicit one (Fahnestock, 2011).

Voice Over: Si Nancy Binay daw 'sing itim ng uling. Pero ang uling 'di ba' pag nag-apoy, anglaking pakinabang sa atin?
(Advertisement #24: Nancy Binay 2019)

The excerpt above shows the political campaign ad of senatorial candidate Nancy Binay. It quoted, *"Si Nancy Binay daw 'sing itim ng uling. Pero ang uling 'di ba' pag nag-apoy, ang laking pakinabang sa atin?" (Others describe Nancy Binay as black as charcoal. But* *once the charcoal burns, 'isn't it beneficial to us?).* Simile was integrated into this ad as a tool of persuasion. A simile is an aesthetic and skilled method of discourse in political speeches whose primary goals are to clarify an opinion or feeling, bring two meanings closer together, and compare one entity to another in praise, disapproval, adornment, or repugnance (Alruzzia & Yunusb, 2019). The result above shows how simile helps the ad clarify the negative connotation about the politician and take this connotation positively.

b. Hyperbole

Hyperbole is a figure of speech that emphasizes a quality or feature by using great exaggeration. Hyperboles are a typical approach used in advertising to grab attention, inject humor into combative dialogue, and accentuate emotions or product features (Barbu-Kleitsch, 2015).

Participant 1: Namasukan ako bilang OFW
Participant 2: Nagkaroon ako ng isang abusadongamo
Participant 3: Naging impyerno ang buhay ko
(Advertisement #28: Jinggoy Estrada 2019)

As shown in the excerpt above, senatorial candidate Jinggoy Estrada's political campaign ad used hyperbole to emphasize the life of being an Overseas Filipino Worker (OFW). In this scene, the OFWs talked about their experiences quoted "*Namasukan ako bilang OFW*" (*I applied as OFW)* said by the first speaker, "*Nagkaroon ako ng isang abusadong amo*" (*I had an abusive boss)* said by the second speaker, "*Naging impyerno ang buhay ko (My life became a hell)* said by the third speaker. In the sentence "*Naging impyerno ang buhay ko (My life became a hell),* the advertisement used the exaggerated word "*hell"* in describing the life of the OFW.

vi. Repetition

Repetition is a literary method in which a word or phrase is used two or more times in a speech or written work for effect. The words or phrases should be repeated within proximity to each other for repetition to be visible. In a literary work of poetry or prose, repeating the exact words or phrases can help to clarify an idea and make it memorable to the reader. In advertising, repetition can help to create a vibrant, charming, and humorous picture (Ming-xin, 2006).

Doon ako sa mabait
Doon ako sa
magalingJV Ejercito
JV is the good one, JV is the good
oneJV Ejercito
JV good, JV good
(Advertisement #9: JV Ejercito 2019)

In Philippine political advertising, repetition emphasized the politician's name as shown in the excerpt above. the candidate's name by using the adjective "good" in *"JV is the good one"* that repeated twice in each line. The results above show that repetition was used to positively and beautifully describe the political actors. It strongly emphasized their names in the ads.

vii. Ad Hominem

According to Walton (2000), an ad hominem argument is the use of personal attack in a dialogue exchange between two parties, in which one party criticizes the other party's character as being bad in some way and then uses this assault as a basis for criticizing the other party's argument.

Robin Padilla: *Simple lang ang kailangan natin, disiplina. Mga train na tumatakbo sa tamang oras,mga kalyeng tama ang pagkagawa imbes na binabaha, mga kalsadang walang traffic, mga airport na hindi kahiya-hiya, pulis na di kriminal bagkus nanghuhuli ng kriminal, mga public servant na hindi nang-aapi kundi nagsisilbi, mga mambabatas na di lumalabag kundi nagpapatupadng batas. Pangulong hindi nagtatago kung hindi namumuno, gobyernong hindi kurap at mamamayang hindi takot palagi. Lahat po iyan pwede pag may disiplina. Aanhin natin ang tuwid na daan kung tinutuntun lamang ng ilan. Tigilan na ang kalokohan.*
(Advertisement #35: Rodrigo Duterte 2016)

In the political campaign ad of presidential candidate Rodrigo Duturte, ad hominem was inserted at the last part of the argument, as shown in the excerpt above. This argument is *"Aanhin natin ang tuwid na daan kung tinutuntun lamang ng ilan"*. The line "*tuwid na daan*" was popularly known in the tagline of the political campaign ads of the former Philippine president Benigno Aquino III including his political party (liberal party). Presidential candidate Rodrigo Duterte's political campaign ad, used this argument negatively and made his campaign platform advocacy shine. The ad emphasized the importance of discipline rather than the "*tuwid na daan."*

Capitals in the Ads

The second stage of CDA, interpretation, is involved with the participants' text production and interpretation processes. The stage of interpretation corrects subjects' delusions of autonomy in discourse as it makes clear what is usually implicit for participants (Fairclough, 2001). Bourdieu's (1986) theory of capital was integrated into this level of text analysis. This study critically analyzed the capitals in the select political campaign ads, as shown in the discussion below.

i. Economic Capital

Economic capital refers to the form of capital that can be converted into money and can be institutionalized in the form of property (Bourdieu, 1986). Unfortunately, this form of capital was not used both in a senatorial and presidential political campaign advertisement. However, results proved that the two remaining forms of capital were mostly used in the senatorial and presidential ads. The discussion of the results is shown in the following subsections of this paper below.

i. Social Capital

Candidates who used social capital in their ads are Tito Sotto, Ralph Recto, Miguel Zubiri, Sonny Angara, Chiz Escudero, Win Gatchalian, Gringo Honasan, Grace Poe, Kiko Pangilinan, Koko Pimentel, Bong Go, Jinngoy Estrada, Bato Dela Rosa, Joel Villanueva, and Rodrigo Duterte (See Appendix B).

Dolphy: Una ka siyang nakilala bilang bata na kalong ng kanyang ama. Paglipas ng panahon, naging kaibigan ko siyang tapat dahil sa kanyangmagandang pag-uugali lalo na sa mga taong pangkaraniwan. Magkatulad kaming nag-umpisasa wala. Ang tanging karangalan naming ang magpasaya at magbigay inspirasyon sa inyong lahat. Noon, ngayon, bukas, at magpakailanman, kayo ang bida. Sa ikasampo ng Mayo, ipanalo natin si FPJ sa pagkapangulo!
(Advertisement #40: Fernando Poe Jr. 2004)

The excerpt above shows how the celebrity acted as the social capital used by the political actor in the ads. The presidential aspirant Fernando Poe Jr was endorsed by Dolphy, an actor and a comedian in the Philippines. The comedian said in the ad that they became good friends because of his good attitude towards ordinary people.

ii. Cultural Capital

Cultural capital can take the following forms: embodied, which refers to long-term mental and physical dispositions; objectified, which refers to the objectification of cultural capital in the form of cultural goods (photographs, books, dictionaries, instruments, machines, and so on); and institutionalized, which pertains towards the objectification of cultural capital in the form of academic or educational qualifications (Bourdieu, 1986).

Mrs. Josephine R. Dela Cruz: Salamat Migz, dahilsa inyong Rent Control Act, hindi na tumaas ang aming upa sa bahay.
Engr. Reden Rodriguez: Thank you Migz, dahil saBiofuels Act of 2006, hindi ko na kailangan mag- abroad para magtrabaho.
Extra: Tumaas ang sahod ko dahil AFP Rate BasePay Act! Thank you, Sir!
Dr. Nielsen B. Donato: Thank you Migz sa WildlifeConservation Act, pinangalagaan mo ang Kalikasan. Champion ka talaga!
Miguel Zubiri: Hangad ko ang kabutihan ng ating mga mamamayan tuwing gagawa ako ng

bagong batas. Tulungan niyo ako sa mga nasimulan ko na.

Ako po si Migz Zubiri!

(Advertisement #3: Miguel Zubiri 2007)

The excerpt above shows the political advertisement of senatorial candidate Migz Zubiri. Juan Miguel "Migz" Fernandez Zubiri was the Former Senate Majority Leader, Former Chairman, Committee on Environment and Natural Resources; Committee on Cooperatives, and Former Chairman, Joint Congressional Oversight Committee on the Ecological Solid Waste Management Act; Joint Congressional Committee on Clean Air Act; Joint Congressional Oversight Committee on Cooperatives. He graduated from the University of the Philippines Los Baños with the degree Bachelor of Science in Agri-Business Management. He took his Master's in Environment and Natural Resources Management at the University of the Philippines Open University (Macaraig, 2013). In his ad, he showcased his cultural capital (in an institutionalized state) by highlighting the bills he wrote and sponsored.

It includes *the Rent Control Act of 2009* (principal author, co-sponsor). We see this in the statement "*Salamat Migz, dahil sa inyong Rent Control Act, hindi na tumaas ang aming upa sa bahay" (Thank you, Migz, because of your Rent Control Act, our house's rental fee doesn't increase,"* said by Mrs. Josephine R. Dela Cruz, *Biofuels Act of 2006* (author) in the statement *"Thank you Migz, dahil sa Biofuels Act of 2006, hindi ko na kailangan mag-abroad para magtrabaho" (Thank you Migz, because of Biofuels Act of 2006, I don't need to go abroad just to find work)* said by Engr. Reden Rodriguez, *Armed Forces of the Philippines Rate Pay Base Increase* (principal author) in the statement "*Tumaas ang sahod ko dahil AFP Rate Base Pay Act! Thank you, Sir!" (My salary got increased because of AFP Rate Base Pay Act! Thank you, Sir!)* said by the participant, Wildlife *Conservation and Protection Act* (principal author) in the statement *"Thank you Migz sa and Wildlife Conservation Act, pinangalagaan mo ang Kalikasan. Champion ka talaga! (Thank you, Migz for Wildlife Conservation Act, you care on the Environment. You are the champion!),* said by Dr. Nielsen B. Donato.

Representations

The goal of the explanation stage is to portray a discourse as a social activity, a social practice, revealing how social structures shape it and what reproductive consequences discourses can have on those structures through time, either sustaining or changing them. In the last stage of CDA, this study focused on the portrayal or representation of politicians' positionality and the social classes position in the select political campaign ads.

A. Representation of Politician's Positionality

Misawa (2010) emphasizes the fluid and relational qualities of social identity formation while also noting that "all parts of our identities are shaped by socially constructed positions and memberships to which we belong" and which are "embedded in our society as a system." In their political campaign ads, politicians' positionality is coated with different ideologies and ideas. Politicians represent their positionalities in the ads in different ways. The following are some identified representations of politicians' positionality.

i. Politicians as our only hope

It is noticeable how "hope" has become one of the representations of the politician's position in the ad. Hope is a strong word that describes how important the person is in the society.

Siya lang ang tunay na pag-asa
Tanging si Loren lang (una sa tiwala)

(Advertisement #15: Loren Legarda 2007)

The example above was extracted from senatorial candidate Loren Legarda, it characterized and represented the politician's positionality as society's only hope. It was observed in the ad of senatorial candidate Loren Legarda the use of the word "pag-asa" ("*hope"*) in representing her positionality in the ad. These lines are “Siya lang ang tunay na pag-asa, tanging si Loren lang” (“*She is the only hope, Loren is the only one*”). This result supports the study of Masroor et al. (2018) that another expression of the positive self-used by political actors involves representing the self or one’s political party as the last hope for the country.

ii. Politician as a Savior

Meanwhile, some candidate represents their positionality in the ad as a savior. In the book of Western (2011) entitled “An overview of the leadership discourses,” he described the term leader as messiah or savior. Messiah discourse provides charismatic leadership and vision in the face of a tumultuous and unpredictable world. The tension between salvation and destruction, between the technocrat and the moral visionary, and between hope and despair is symbolized by the messiah character. The messiah discourse appeals to both individuals and society, promising deliverance from a chaotic world marked by a lack of control and a diminishing sense of community (Western, 2011).

Participant 1: Bilisan
mo!
Participant 2: Nariyan na
sila! Participant 1: Dali! Ang
Bato! Participant 2: Bato
Bato sa langit!
Voice Over: Dahil kay Bato Dela Rosa,
bumababaang krimen. Nabago ang buhay ng
mga nasa mundo ng droga. Kalaban ng
terorismo. (Advertisement #29: Bato Dela
Rosa 2019)

As shown in the excerpt above, the ad contains a dramatic dialogue. The concept of the ad was adopted from the scenes of *Darna*, one of the most iconic superheroines in the Philippines. *Darna* was the brainchild of komiks (Filipino colloquial for comic books) written by Mars Ravelo (Llanes, 2009). The candidate’s nickname “Bato” was highlighted in this ad as a magic white stone that saves and protects the people from evildoers. “Bato,” a Filipino term for “stone,” was very popular in *Darna* because when *Narda*, the main character of the story, swallows a magic white stone, she transforms into the mighty warrior *Darna* by shouting "DARNA!". This dramatic dialogue was seen in the lines “Bilisan mo” (*“Make it fast”)* said by participant 1, “Nariyan na sila” (*“They are here”)* said by participant 2, “Dali! Ang Bato” (*“Faster! The stone!)* said by participant 1, and “Bato Bato sa langit” (*“Stone Stone in the sky”)* said by participant 2 followed by a voice over saying “Dahil kay Bato Dela Rosa, bumababa ang krimen. Nabago ang buhay ng mga nasa mundo ng droga. Kalaban ng terorismo” (*“Because of Bato Dela Rosa, the crime was lessened. The lives of the people involved in drug operation were changed. The enemy of terrorist),* these lines acknowledge the contribution of the political actor in lessening the crime rate in the Philippines.

i. Politician as a person who grew up inpoverty

Understanding hegemony requires Gramsci's view of ideology as an organic relationship between structure and superstructure. According to Gramsci (1978), ideology is conceived as a practice-producing subject. One of the observations of this study is the subject or representation of the politician’s positionality as a person who grew up in poverty.

Manny Pacquiao: Lumaki ako sa hirap. Lahat
ngaking pinagdaanan ay naging dahilan para
ako’ykumalinga at tumulong sa inyo. Manny
Pacquiaopo, nandito para sa inyo.
(Advertisement #16: Manny Pacquiao
2016)

The excerpts above show how the politicians described their positionality in their ads. “Lumaki ako sa hirap. Lahat ng aking pinagdaanan ay naging dahilan para ako’y kumalinga at tumulong sa inyo. Manny Pacquiao po, nandito para sa inyo” (*I grew up in poverty. All that I’ve experienced became the reason to care and to help you. Manny Pacquiao is here for you”.)* senatorial candidate Manny Pacquiao said in his ad. This subject or representation of his positionality in his ad as a person who grew up in poverty formed an ideology that a person needs to look back on where he came from.

A. *Representation of Social Classes'Position*

This section will discuss the representation of social classes' position in the ads to help the current study identify the social dynamics that play a crucial role in examining the hegemonic features of Philippine political advertising. Gramsci (1978) argues that analyzing the historical conditions necessary for one class to acquire hegemony over others. Politicians also considered the background of their audiences who will watch their political campaign advertisements. This study has observed that many of them were careful in representing the social classes in the Philippines. Still, some politicians represented other classes, particularly the subaltern class, in a subjective way.

The representation of SubalternGroup

Representation of how the ads position or describethe social classes helps this study analyze the hegemonic class's social dynamics and dominant ideas or ideologies.

Voice Over: Sino ba si Nognog? Si Nognog ay angbawat Pilipinong bilad sa araw, yung walang matinong classroom, yung napipilitang magtrabaho sa abroad, yung pinagbibintangan sa hindi niya kasalanan, yung walang pambili ng pagkain. Kaya sa mga Nognog, nognog din ang nagmamalasakit. Nognog pero hindi tulad ng iba, marami ng nagawa.
(Advertisement #33: Jejomar Binay 2016)

The excerpt above describes the subaltern class as "nognog" in the ad of presidential candidate Jejomar Binay in "Sino ba si Nognog? Si Nognog ay ang bawat Pilipinong bilad sa araw, yung walang matinong classroom, yung napipilitang magtrabaho sa abroad, yung pinagbibintangan sa hindi niya kasalanan, yung walang pambili ng pagkain. Kaya sa mga Nognog, nognog din ang nagmamalasakit. Nognog pero hindi tulad ng iba, marami ng nagawa" (*Who is Nognog? Nognog" is every Filipino who labors under the sun, who endures studying in makeshift classrooms, the one who is forced to work abroad, the one who is accused of the crime he never committed, and who has barely enough money to eat every day. So, with the Nognog, Nognog also cares. Nognog, but unlike others, he had many accomplishments).* The term "*nognog*" is short for "*sunog*" or burnt, a reference to a dark, short, with curly hair and a stubby-nosed comic character created by L.S. Martinez in the 1970s (Cepada, 2016). This representation for those Filipinos who belong in the subaltern class can relate to this ad because this representation captures the reality of their lives. The politician also positioned himself as one of them but from a different angle, a *Nognog* who cares (referring to the politician).

CONCLUSION

The study's finding is essential in the field of social research. The lack of awareness of the meanings, ideas, and ideologies residing in the discourse of politicians will not help the society or country in achieving social justice, class equality, freedom, rights, and economic growth. Furthermore, increasing the standard for creating political ads in the country will challenge and measure the politicians' communication, competencies, leadership, and excellence. In this way, politicians may change their perspective about politics - that it is not merely the use of power, dominion, or control. Hence, it is all about serving the country with the right motive and motivation. Future researchers who are also interested in conducting a study on Philippine Political Advertising or other existing Political Discourses may include political campaign advertisements during the local election. They may also adopt other tools or methods in the analysis of ads such as multimodal analysis. In addition, future researchers may use other theories in social research and other frameworks in Critical Discourse Analysis.

References

Alruzzia, K. A., & Yunusb, K. B. (2019) Creating Discourse Using Figures of Speech in the Speeches of King Abdullah II. *International Journal of Innovation, Creativity, and Change, 8*(9), 245-253

Balmores-Paulino, R. (2018). An Exploration of the Schema and Function of Humor. *Israeli Journal of Humor Research: An International Journal*, *7*(2), 43-63.

Barbu-Kleitsch, O. (2015, May). Use of hyperboles in advertising effectiveness. In International Conference Redefining Community in International Context (Vol. 15).

Bourdieu, P. (1986). *"The Forms of Capital."* Pp. 241- 58 in Handbook of theory and research for the sociology of education, edited by J.G Richardson. New: Greenwood Press.

Brader, T. (2005). Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. American Journal of Political Science, 49(2), 388-405.

Cepeda, M. (2016, January 13). In the new Binay ad,the average Filipino is 'nognog' like him. *Rappler*. https://www.rappler.com/nation/elections/11 89 20-binay-ad-average-filipino-nognog/

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research.* New York:Routledge.

Fairclough, N. (1995) *Critical Discourse Analysis*.(London, Longman).

Gramsci A (1971) *Selections from the Prison Notebooks (trans. Q Hoare and GN Smith).* New York: International Publishers.

Gramsci, A. (1978). *The Modern Prince and Other Writings.* New York. International Publishers.

Harris, M. K. (2009). The Political application of humor.

Ilhamsyah I. & Herlina H. (2019) *"Tagline in Advertisement Digital Era Case Study of #adaaqua Advertising Campaign."* Neliti. https://www.neliti.com/publications/293392 /ta gline-in-advertisement-digital-era-case-study-of-adaaqua-advertising-campaign#cite

Llanes, R. (2009, July 17). Darna Through Six Decades. *Pep. ph.* https://www.pep.ph/lifestyle/19501/darn a- through-six-decades

Macaraig, A. (2013). This time, Migz Zubiri wants untainted victory. Rappler. https://www.rappler.com/nation/elections/19 73 4-this-time-migz-zubiri-wants-untainted-victory/

Masroor, F., Khan, Q. N., Aib, I., & Ali, Z. (2019). Polarization and ideological weaving in Twitter discourse of politicians. Social media+ society, 5(4), 2056305119891220.

Ming-xin, L. I. U. (2006). Rhyme and repetition in advertising English and translation approaches [J]. Journal of Changsha Telecommunications and Technology Vocational College, 4.

Newsome, C. (2002). "The Use of Slogans in Political Rhetoric." The Corinthian, 4 (3). https://.kb.gcsu.edu/cgi/viewcontent.cgi?arti cle =1203&context=thecorinthian

Rahmawati, L., & Tajuddin, S. (2020). Analysis of Visual and Persuasive Language Used on Billboards Promoting Candidates of the 2019 Elections. KnE Social Sciences, 668-678.

SET.GOV.PH. Republic Act No . 9006.https://www.set.gov.ph/resources/elect ion- law/republic-act-no-9006/

Van Dijk, T. A. (1993). Principles of Critical Discourse Analysis. Discourse and Society 4(2): 249-83.

Vonderau, P., Florin, B., & Zimmermann, Y. (2021). Advertising as Commercial Speech: Truth and Trademarks in Testimonial Advertising. In Advertising and the Transformation of Screen Cultures (pp. 195–214). Amsterdam University Press. https://doi.org/10.2307/j.ctv1t1kgh7.9

Walton, D. (2000). Case study of the use of a circumstantial ad hominem in political argumentation. Philosophy & rhetoric, 33(2), 101-115.

Western, S. (2011). An overview of the leadership discourses. Educational leadership: Context, strategy and collaboration, 11-24

Morphological Paradigm of Nouns and Verbs: Meaning and Functions in Bisakol, a Philippine-type language

Ana Cristina G. Fortes
Sorsogon State University
Sorsogon City, Philippines
anacristinafortes05@gmail.com

Abstract

The linguistic documentation of Philippine-tyoe languages and efforts to revitalize them have increased with the introduction of the MTB-MLE to the primary grades. Language varities in Sorsogon, including the Southern Bisakol, are widely used but less studied by native speakers. Among the grammatical categories investigated are nouns and verbs. They are content words and are usually introduced in the primary grades. Using Payne's (1997) morphosyntax and Nolasco's (2007) stem-based hypothesis, this descriptive study employed structural analysis on the transcribed and compiled Bisakol corpus. Significant findings show that nouns and verbs are derived forms with specific morphemes that may be attached either to the root or to the stem. The affixes of Bisakol have some conditioning and restrictions in the linguistic environment. Hence, Bisakol morphemes, especially their affixes, are highly multifunctional with every affix carrying a meaning that may either modify the semantics of the root or change the concept category of the new form entirely.

1 Introduction

Philippine-type languages (PL) have many complexities and intricacies that entice many scholars of linguistics to investigate. Among the many Philippine-type languages that exhibit diversity and complexity are the Bikol and Visayan languages. Bikol is a macro language that has widely varying dialects and closely related languages. The individual languages that correspond to a macro language are very closely related, and there must be some domain in which only a single language identity is recognized (Lewis, Simons, & Fenning, 2015). In many areas in the region, speech varies dramatically over a few kilometers from one town to the next (Lobel and Tria (2000).

Interesting in Sorsogon, a province in the Bicol region, is the language variety that the speakers themselves believe to have been a mixture of Bikol and Binisaya languages. Natives would name their language Bisakol. Ethnologue names this language as Sorsoganon, Northern; Sorsoganon, Southern; and Masbateno. Earlier linguistic studies particularly, McFarland (1974) and Zorc (1977) confirmed that Southern dialects including the Northern Masbate, Northern Sorsogon and Southern Sorsogon or Gubat belong to the Central Visayan subgroup, of which Hiligaynon and Samar-Leyte are members (p. 299). The geographical location of the province has a major contribution to why these varieties are linguistically Visayan rather than Bikol. The absence of the speaking roads that linked Bikol towns with one another for a long period of time, and the transportation by water brought about this present-day linguistic situation. The proximity of Sorsogon to Samar and Masbate and the water transportation that happened from Samar to Sorsogon, and vice versa, can explain the mixture of Binisaya and Bikol in the dialect in Southern Sorsogon.

At present, there has been a dearth of studies on the languages spoken in Sorsogon area. The differences among the varieties of Bikol Sorsogon are validated in the dialectology of Cunanan (2015). Escalante (1978) and Nolasco (1994) made studies on the grammar of Sorsogonan. Escalante made a description of the internal structures of South Sorsogon verbs and Nolasco worked on the grammar sketch of the language.

2 Morphology of Philippine-type languages: Theoretical Views

Morphological structures of Philippine-type languages are diverse both in their uses and meaning. Segmenting words into meaningful parts, the morphological shapes revealed some systematic covariation in form and meaning (Haspelmath & Sims, 2010). The adjustment in the shapes of the words depends on the way the speakers intend their utterances to be interpreted (Payne, 1997). In Tagalog, for instance, some affixes, when attached to another morphological form, express clearly their language-specific meanings. Philippine-type lexical bases and their meaning change via affixation (Himmelmann, 2008). The lexical bases can be syntactically subcategorized as content words, such as nouns and verbs. Classifying a root or a base either a noun or a verb in Philippine-type languages appears to be challenging. The concept of precategoriality indicates the idea that for any root to be classified as either noun or verb, it should be affixed or case marked. Foley (1998) expressed the lack of noun-verb distinction of roots in Philippine languages.

To some scholars, nouns and verbs in Philippine-type languages have language-specific features. For Payne (1997, 2021), to determine the grammatical category of a given lexicon, the linguist needs to identify the morphosyntactic characteristics of the prototype. Payne proposed the morphosyntactic analysis with its two important properties: distributional and structural properties. "Distributional properties show how words are distributed in phrases, and clauses and texts. Structural properties illustrate the internal structure of the word itself" (p.33).

The morphological process particularly affixation is highly noticeable in Philippine-type languages. The stem-based hypothesis introduced by Nolasco (2007) demonstrates the layering of morphemes in nouns and verbs in Tagalog. This method of analysis clearly shows that word forms in Tagalog and other Philippine-type languages have similar meanings since they are linked to one stem, although the word forms may differ in the arguments that they co-index in the clause construction.

Bisakol, a Philippine-type language demonstrates the same morphological features as that of Tagalog. By morphological form alone, nouns and verbs in Bisakol may demonstrate the same features. However, building the analysis on Payne's (1997) morphosyntactic operations in determining the meaning and functions reveals the language-specific features of nouns and verbs in this language. Bisakol words are formed by several morphological processes such as affixation and reduplication and these word forms have multiple affixes layered in their structure.

3 Purpose of the Study

The purpose of the study is to (1) analyze and describe the morphological features of nouns and verbs in Bisakol along with the different morphological processes in the language, and (2) establish a morphological paradigm that would guide mother tongue teachers in Bisakol in designing their learning content in mother tongue-based teaching.

4 Methodology

This study is a typological investigation of the linguistic features of the language variety in Southern Sorsogon. The study analyzes the morphological structures of the nouns and verbs through the morphosyntax framework of Payne (1997) and the stem-based hypothesis' of Nolasco (2007).

As the researcher has been building corpus of Bisakol from spoken and written text, this study made use of the existing eighty-two thousand corpus collected since 2016. The spoken texts are stories of the informants, conversations among college students and family members, and public speeches made by candidates during the 2016 national election. The oral texts were recorded, transcribed, and secured in an electronic file. Likewise, the online discourse particularly those found in the online group pages and accounts of some Bisakol speakers are also added to the collected Bisakol corpus. These social media group pages are accessible online. Another data collection procedure used by the researcher is wordlist and elicitation. Using the word list provided by Reid (2016), two informants provided the word list equivalent in Bisakol. In addition, using some sentences in Tagalog, the researcher requested two informants to translate each into Bisakol.

5 Analysis and Findings

Affixation and reduplication are morphological processes in Bisakol that allow for a word to be formed. Despite some claims of scholars on the lack of noun-verb distinction (Foley, 1998) of roots in Philippine-type languages, findings of this study reveal that nouns and verbs are distinct categories. The morphological analysis reveals that nouns are formed by nominalizing affixes and verbs are affixed for voice, aspect, and mode. Affixation and reduplication, the morphological processes, are operationalized by the syntactic conditions or restrictions in the language.

5.1 Morphological Features of Bisakol Nouns

In Bisakol, nouns become complex when it is composed of the root or stem and an affix. Complex forms of nouns are those that undergo a certain morphological process and change the word class, say verbs to nouns. Commonly, they are derived forms or nominalized forms. Nominalization is a morphological derivation that creates or forms a noun stem. The new stem derived from verbs or adjectives belongs to the noun class (Payne, 2006).

Bisakol nominalizing affixes are grouped according to their morphological behavior. First, the prefixes *pag-, paN-, para-, ka-, taga-, tig-, tag-* are attached to Bisakol root or stem to form words. Second, the fused affixes namely *pag-+pa-, pag-+ka-, paN-+pa-* are placed before the root or stem. Interestingly, the affixation of *pa-* and *ka-* is permitted to be fused when the morphemes *pag-*, and *paN-* are added to the root or stem. This process is called morphological layering. It is a process in which an affix is required to be attached to the stem first before another affix can be added. Third, Bisakol has circumfix, *ka-...-an* in which the *ka-* is added before the root and *–an* is added at the end of the root. Lastly, voice nominalizing affixes are attached to Bisakol root or stem. They are the null *Ø (where m- / -um are replacive affixes), (h)-an, i-, -un,* and *–in* (see Table 1). Illustrating the morphological structure of nominalized forms, some examples of Bisakol nouns are shown in (1)-(4).

(1)pag-$_{\text{[affix]}}$ + kaon$_{\text{[root]}}$ *'eat'* > pagkaon *'food'*

(2) pag-$_{\text{[affix/layer1]}}$+pa-$_{\text{[affix/layer2]}}$+ daku $_{\text{[root]}}$ *'big'* > pagdaku *'growth'* > pagpadaku *'to raise a child'*

(3)ka-$_{\text{[affix]}}$+tuyu$_{\text{[root]}}$*'intent'*+(h)an$_{\text{[affix]}}$ > katuyuhan *'intention'*

(4) bayle$_{\text{[root]}}$*'dance'*+-(h)an$_{\text{[affix]}}$ > *baylehan* 'a place to dance'

Clauses (5) and (6) are presented to show the nominalized forms in Bisakol syntax. In (5) and (6), the nominalized forms are *pagkakonsehal* and *pag-eskwela*. The nominal function of these two forms is case marked by the determiners **sa** in (5) and **an** in (6). Absolute determiner, ***an*** marks nouns and noun phrases (NP) in Bisakol. The determiner ***sa*** in Bisakol marks spatial location, temporal location, a nominal expressing goal as its macrorole. Similarly, nominal forms in Tagalog and other Philippine-type languages are case-marked by absolutive determiners (Dita, 2010b) or by locative determiner.

Affixes	Type	Position
pag-; paN-;para-; ka-; taga-; tig-; tag-	derivational	Prefix
pag-+pa-;pag-+ka; paN-+pa	derivational	Prefix
ka…an	derivational	circumfix
Ø(m-/-um); in-/-in	derivational	prefix/ infix
-an; -un	derivational	suffix
-i	derivational	prefix

Table 1: Affixes in Bisakol Nouns

(5) ...nagdalagan sa pagkakonsehal si Papa.
nag- dalagan sa **pag-ka- konsehal** si Papa
ASP- run LOC NMZLR-ABL councilor ABS father
'[My] father ran for [municipal] councilor.'

(6) Mahal an **pag-eskwela** sa Manila
mahal an **pag-** eskwela sa Manila
expensive ABS NMZLR- schooling LOC Manila
'It is expensive to study in Manila.'

5.2 Morphological Features of Bisakol Verbs: Aspect, Voice, and Mode Affixes

The morphology of verbs in Bisakol is complex because of the of the considerable morphophonemic fusion and alternations of affixes in the language. The verbs are affixed with morphemes that inflect for voice, aspect, and mode. Thus, verbal affixes are grouped according to their coded syntactic functions. The first group of affixes that inflect for voice *include i-, -an, -in, -a, Ø>um/~m* (replacive affix *m-/n-).* The second group of affixes inflects for aspect. Aspectual affixes include the *mag-, nag-* and *in-/<in>*. The mode affixes are *pag-,-ang, ka-, pa-, paki-, <V1r>*.

Voice Affixes: For Cena (2014) voice refers to the "affix in the verb that imposes a thematic role reading on the subject which refers to the absolutive nominal" (p.200). Putting it simply, the manifestation of voice is found in the affix attached to the verb in which the affix co-indexed a noun as the most affected entity of the action expressed by the verb or the most agentive or the instigator of the action. To present the morphological structure of verbs inflected with voice affixes, examples (7)-(10) are provided. The actor, patient, and locative voice affixes are attached after the root. The alternative morphological structure is shown in (10) or the instrument voice. The affix ***i-*** can be attached before the root or after the root.

Verbs indicate an actor voice (AV) with the null affix (Ø) attached to the them. In cases when the verb changes aspect, the replacive affixes *m-/n-* appear in the stem. The argument bears the actor macro role or the undergoer of the process of the action. The patient voice with the *-un* affix co-indexes the patient or goal. The locative voice in Bisakol is marked by the affix *-an* attached to the verb. This affix co-indexes an argument which may express semantic roles such as, location, recipient, benefactive, or goal. When a verb is voiced-mark by *i-* affix, the action is directed towards the argument which may have the semantic roles of the instrument or the transported theme (Tanangkingsing, 2009).

(7) actor voice: Ø + root

Ø~mag-$_{\text{[affix]}}$+hatag$_{\text{[root]}}$ *'give'*>maghatag *'give something'*

Ø~nag[affix] +kadi[root]'come'> nagkadi 'came'

(8) patient voice: root + -un

hatag[root]'give' + -un[affix] > hatagon 'give something to someone'

sabut[root] 'understand' + -un[affix] > sabuton 'make one understand something'

(9) locative voice: root + -an ~-a

hatag[root] 'give' +- an[affix] > hatagan'give someone something

himu [root] 'give' + -an [affix] > himuan 'make someone something'

(10) instrument voice: i- + root

i-[affix] + hatag [root] 'give' > ihatag 'give [directed someone] something to someone'

root + -i

butang[root] 'put' + -i [affix] > butangi 'put/pour something onto something'

Clauses (11), (12), and (13) illustrate the illustrate the morphosyntax of the voice affixes in the language. The clauses show the transitive construction with two core arguments marked by the verb ***hatag*** *'give'*. When affix ***-an*** is inflected, locative voice is expressed. The argument, ***an phra*****se**, is the recipient of the verb which semantically implies morphosyntax of the voice affixes in the language. The clauses show the transitive construction with two core arguments marked by the verb hatag 'give'. When affix -an is inflected, locative voice is expressed. The argument, an phrase, is the recipient of the verb which semantically implies that there is a movement of the object from someone to another one (11). Meanwhile, the ***i-*** affix implies instrument voice (12), and affix ***-un*** marks the two core arguments ***mu*** and the ***an-*****phrase** in which the *allowance* is the transported theme (13).

(11) Hatagan ku si mama kwarta.
hatag-an= ku si mama Ø kwarta
give-VOICE.LV=ERG.1s ABS mother OBL money
'I will give [my] mother [some] money.'

(12) Ihatag mu an kwarta kay mabayad aku sa eskwelahan.
i- hatag=mu an kwarta kay
VOICE.IV-give =ERG.2s ABS money because
ma- bayad=aku sa eskewelahan
ASP- pay =ABS.1s LOC school
'Give me that money because I have to pay [something] in the school.

(13) Hatagun mu na lang an allowance kaniya.
hatag-un =mu= na=lang an
give-VOICE.PV=ERG.2s=PAR=PAR ABS
allowance kaniya
allowance OBL.3s
'[Just] Give her the allowance.'

Aspectual Affixes: Aspect means the different ways of viewing the internal temporal constituency of the situation (Holt, 1943 in Comrie, 1976). Unlike tense which relates to the time of situation referred either to moment of speaking or to some other time, aspect is concerned with the features of completion, durability, perfectivity, or imperfectivity of the action or process. In Philippine languages, aspect is highly referred than tense. Speakers of PLs, particularly, of Bisakol, are not mainly concerned with the temporal relation of the action relative to the moment of speaking but rather with the internal temporal constituency of one situation or action. The situation of internal time shows whether the action started yet unfinished; it started and finished, or haven't started yet. The six aspects of Bisakol are expressed morphologically. Aspectual affixes include the *ma-, mag-, na-, nag-, and in-/<in>*. Reduplication is highly used in verbs to express aspect. Morphological structure of each aspect is shown in (14) to (19).

(14) neutral/infinitive$_{\text{[-beg][-fin]}}$:
INTR: mag + root
mag-$_{\text{[affix]}}$ + surat$_{\text{[root]}}$ *'write'* > magsurat *'to write'*
< ~um>+ root
s<um>[affix]urat[root] > sumurat *'to write'*

(15) perfective aspect $_{\text{[+beg],[+fin]}}$
INTR: nag- + root
nag[affix] + surat[root] 'write' > nagsurat 'wrote'
TR: in- + root
in-[affix] + bakal[root] 'buy' > inbakal 'bought'
TR: <in> + root
h<in>[affix] atag[root] 'give'> hinatag 'gave'

(16) Recent Perfective [+beg][+fin]
INTR: na-+ka-+ root
na-[affix]+ ka-[affix] + surat[root] 'write' :
nakasurat 'be able to write'
TR: ka- + (:) + {C1V1} [redup] + root + -a
ka:[affix][vowel leng] + {ba} + bakal [root]
'buy'+ -a > kababakala 'has just bought'

(17) Actual Imperfective [+beg],[-fin]
INTR: nag- + {C1V1}[redup]+ root
nag-[affix]+ {su}+ surat[root] 'write' >
nagsusurat 'is writing'
TR: -in-+ {C1V1}[redup] + root
in-[affix] + {ba}+ bakal [root] 'buy' > inbabakal
' is buying'

(18) Conditional Imperfective [+beg],[-fin]
INTR: na-+ : + root
na: [affix][vowel leng]+ surat [root] 'write' >
na:surat 'have written'
TR: (k)i +: + root + -an
ki:[affix][vowel leng]+ kaun[root] 'eat'+ -
an[affix] > ki:kaunan 'have eaten'

(19) Contemplative [-beg], [-fin]
INTR: ma: + root
ma:[affix][vowel leng] + imod[root] 'see' >
ma:imod 'will be watching'
INTR: mag- + {C1V1}[redup]+ root
mag-[affix] +{su}+ surat[root] 'write'>
magsusurat 'will be writing'
TR: {C1V1}[redup]+ root+ -un/an
{ba}+ bakal[root] 'buy' + -un/an > babakaun /
babaklan 'will be buying'
TR: i- + {C1V1}[redup]+ root
i-[affix] + {ha} + hatag[root] 'give'> ihahatag
'will be giving'

The verbs in the neutral aspect are noted for an action has not started, therefore, no completion can occur. They are often referred to as infinitive forms. Perfective aspect indicates a completed action. For recent perfective, it denotes an action that has just been done or completed recently. In Bisakol, the affix ka- is added to the stem and the reduplication of C_1V_1 of the base. Actual imperfective aspect expresses an action that has begun but has not been completed. The conditional imperfective aspect expresses an idea that a certain action is performed when the person is in certain conditions or usual scenarios. Contemplative aspect refers to the action or state that has not started yet.

Affixes in Bisakol that carries the grammatical voice like *-un* and the aspectual affix *-in* are in complementary distribution. When the affix *-in* is added to the verb, the voice affix *-un* becomes null. The *in- /<in>* when inflected to verbs in the patient voice is sufficient to mark two important features of verbs in Bisakol, the voice and aspect. Guzman (1994) explains this phenomenon as the principle of minimal distinction or the inflected forms function identically in syntax even if the morphological paradigms are regular or irregular. Clauses (22) and (23) demonstrate this morphological feature.

(20) root + -un :
surat$_{\text{[root]}}$'write'+-un $_{\text{[affix] >}}$suraton 'to write'
(21) -in + root + -un ~ Ø
-in$_{\text{[affix]}}$ + surat $_{\text{[root]}}$ + Ø$_{\text{[affix]}}$ > insurat 'wrote'

(22) Suratun mu an ngaran niya sa papel.
surat-un=mu an ngaran=niya sa papel
write-PV=ERG.2s ABS name =GEN.3s LOC paper
'You write his name on the paper.'

(23) Insurat mu an ngaran niya sa papel?
in- surat-Ø= mu an ngaran=niya
ASP-write-PV=ERG.2s ABS name=GEN.3s
sa papel
LOC paper
'Did you write his name on the paper?'

Mode Affixes: Mode is one one feature of verbs in Bisakol. It relates to the manner of realization of the action. Two verbs may express similarity in aspect, and even voice, but their difference in mode is recognizable as a language-specific feature of Bisakol verbs. The imperative mode indicates command. The *-ang* mode affix expresses either extensiveness and intensiveness of how the action is performed by the agent or actor. The *ka-* affix expresses abilitative mode. It refers to the ability of actor or agent to perform the action indicated by the verb. However, *ka-* expresses another mode of the verb in Bisakol such as motive, unexpected result and reason. Causative mode indicates that the actor or agent has caused the action to be done. The requestive mode expresses social, commitative, and permissive sense. The distributive mode expresses an action that has been participated by all actors mentioned in the clause. The morphological structure of verbs expressing different modalities are shown in examples (24) to (30).

(24) imperative
pag- + root
pag-$_{[affix]}$+ kaon$_{[root]}$ > **pag**kaon (na) ‘you have to eat’
root + **-i**
hatag[root] + -i[affix] > hatagi ‘give someone something’
root + **-a**
hatag[root] + -a[affix] > hataga ‘give something to someone’

(25) extensive/intensive
m-/p- + **-ang** + root
m-$_{[affix]}$+ -ang$_{[affix]}$+ limpya $_{[root]}$ > m**ang**limpya ‘will clean extensively’
p-$_{[affix]}$ + -ang$_{[affix]}$ +limpya $_{[root]}$ > p**ang**limpya ‘to clean extensively’

(26) abilitative/aptative
m-/p- + **ka**- + root
ma-$_{[affix]}$ + ka-$_{[affix]}$+ pasar$_{[root]}$ > ma**ka**pasar ‘to pass’
pa-$_{[affix]}$+ ka-$_{[affix]}$+ pasar$_{[root]}$ > pa**ka**pasa ‘able to pass’

(27) causative
nag-/mag-/pag- + **pa**- + root
nag-$_{[affix]}$+pa-$_{[affix]}$+kaon$_{[root]}$ > nagpakaon ‘feed someone/something’

(28) requestive/ commitative
maki-/paki + root
maki$_{-[affix]}$ + huron $_{[root]}$ ‘discuss’ > makihuro ‘discuss something about’

(29) Motive/unexpected
mag-/pag- + **ka-**+ root
nag-$_{[affix]}$ + ka-$_{[affix]}$+ hapdos$_{[root]}$ ‘illness’> nag**ka**hapdos ‘got sick’

(30) Distributive/plurative
root + { $\mathbf{V_1r}$**-**}
in$_{[affix]}$+k-+{ar}on$_{[root]}$ + -an$_{[affix]}$ > inkaraunan ‘have something used for eating’

5.3 Layering of Affixes in Bisakol Verbs

Illustrating stem-based hypothesis to Bisakol morphology, word forms in Bisakol are multi-layered. An affix is added to the root to form a stem which prepares the word form for another affixation process. The layering of affixes is shown in example (31). The root *surat* ‘write’ is affixed with the nominalizing affix, *pag-*. The word form *pagsurat* ‘to write’ may express an infinitive aspect or imperative mode. When the *p-* is replaced with n-, the word form is *nagsurat* ‘wrote’. The nag- is an aspectual affix expressing perfective aspect. When the first syllable of the root is reduplicated and added to the stem, *nagsurat* ‘wrote’, the word form becomes *nagsusurat* ‘is writing’. The reduplication C_1V_1 expresses imperfective aspect. In the example, the final word that may be formed is *nagpasurat* ‘to have someone write something’. The word formed has several layered affixes such as nag- is an aspectual affix, *pa-* is a mode affix, reduplicated *–su*, express imperfective aspect, and *pag-* is a nominalizing affix. Hence, this layering of morphemes reveals that Bisakol is a highly inflectional language in which affixes often fuse to code several grammatical functions and meanings.

(31) surat[root] ‘write’
>pag-[nom.affix] surat [root] ‘to write’
> nag-[asp.affix] surat[root] ‘wrote’
>nag-[asp.affix]{su}[redup.asp] + surat[root] ‘is writing’
> nag-[asp.affix]pa-[mode.affix]surat [root] ‘has asked something to write something’

5.4 Multifunctionality of Bisakol Affixes

From the analysis, the study was able to determine that Bisakol affixes are highly multifunctional. Affixes in verbs expressing voice are also used as nominalizing affixes such as *–an, -un, i-,* and the null affix for intransitive construction. When these voice affixes are added to the root or stem, these affixes shape the meaning of the nominal forms to express the concept category associated to the root. By affix *-an*, the stem becomes a nominal form of a place of location and *–un* expresses a trait or a state or condition. The claim that these forms are nominalized forms grammatically functioning as nouns is asserted by the presence of absolutive case markers in Bisakol.

(32) **an** laba-{h}**an** ‘a place to do laundry’
ABS wash –NMLZR

(33) **an** sugna **-an** 'the place to do cooking'
ABS cook- NMLZR
(34)**sa** higda-**an** 'the place to lay down'
LOC lay NMLZR

In analyzing verbal affixes particularly those expressing mode, the affixes *pag-* and *ka-* are mode affixes, yet, nominalizing affixes too. The affix *pag-* expresses an abstract idea or a concept when used as nominalizer. The affix *ka-* expresses a commitative of partative entity in Bisakol nouns. However, *pag-* is a mode affix expressing imperative construction when added to the verb. The *ka-* is a mode affix expressing abilitative or aptative, motive, or unexpected occurrence.

(35) pag- [nom. affix] vs pag- [verb affix]
(35.1) pag- tubod>pagtubod 'faith'
NMLZR believe
(35.2) Pagkarigos na.
pag- karigos =na.
IMP- take a bath PAR
'You take a bath now.'

(36) ka- [nom. affix] vs. ka [verb affix]
(36.1) ka- upod> kaupod 'company'
NMLZR join with
(36.2) Gustuhon ku man makapasar sa exam ni sir.
Gustu-(h)un=ku =man ma- ka- pasar
Like -INTSFR=ERG.1s =PAR INF- ABL-pass
sa exam ni Sir
LOC exam GEN Sir
'I really would like to pass in the exam [of Sir]'

Significantly, the mutual exclusivity of the affixes –*un* and –*in<in>* in Bisakol verbs reveals that an *affix* can carry two grammatical functions in the clause. The *in- /<in>* when used with patient voice is sufficient to mark two important features of verbs in Bisakol, the voice and aspect (see sample 22-23).

5.5 Morphological Paradigms in Bisakol

Stressing the argument of pre-categoriality or the lack or noun-verb distinction in Philippine-type languages, this study strongly states that nouns and verbs in Bisakol are two separate and distinct word classes. Nominalization occurs because the language has various morphemes that derive nouns from another word class, and so with verbalization. Nouns are case-marked by the determiners of the language such as *an, si, ni, san/sin,* or *sa* and these nouns fill the argument slots in the clause. Verbs in Bisakol are inflected for voice, aspect, and mode. Verbs do not fill any argument slot but index the number of arguments that are required to be present in the syntax of the language. Reduplication is another morphological feature, especially in verbs. The reduplication marks a grammatical function, particularly in the aspect of verbs. Verbs in Bisakol are affixed to indicate how native speakers view time, manner, and transitivity in their language.

Grammatical Category	Morphological	
	Feature	Processes
Nouns	free and bound root and stem	Affixation (nominalizing affixes) Reduplication
Verbs	stems are highly layered with affixes inflecting for aspect, tense, and mode	Affixation (aspect, voice, mode) Reduplication

Table 2: Morphological Paradigm of Nouns and Verbs

6 Conclusion

This study presents the morphological characteristics of nouns and verbs in Bisakol. Bisakol affixes are highly multifunctional. Nouns are mostly derived forms. The affixes change the word forms into nominalized forms making them occupy argument slots in the clause. Verbs are a distinct category because they can be affixed only for voice, aspect, and mode. Significantly, the nominal and verbal affixes when attached to the root or stem create different concept categories. Word forms, especially verb forms are layered with several morphemes that code for different grammatical meanings and functions. With morphophonology, the affixes of Bisakol have some conditioning and restrictions, for instance, the affix *-in/ <in>*, although an aspectual affix, marks a patient voice too. Aspect, voice, and mode are verb features that are marked by affixes in which only Bisakol speakers and their intuition in their native language can capture the appropriate contextual meaning.

7 Implications to MTB-MLE in the Sorsogon

The scenario of the teaching of the mother tongue in the classrooms and the use of mother tongue in teaching other disciplines has been challenged by so many compelling factors. Among the many are the teachers' knowledge and competence in their own L1 which is also the L1 of the learners. This study strongly considers that teacher's understanding of the grammar of language that they speak highly contributes to the success of the teaching-learning process and preparation of instructional materials in the mother tongue. By being equipped with the concept of how language works in L1, teachers can decipher how their learners cognitively and metacognitively process information and develop grammatical concepts in L1 and L2. Thus, teachers could properly introduce the concept of time in L1 which is more on aspect and the concept of time in English which is viewed more as a tense. Beyond differentiation of grammatical concepts, L1 teachers who will become aware of the morphological processes in the language can help learners build their L1 vocabulary, develop phonemic awareness, and phonic skills. The common underlying proficiency in L1 and L2 can be grounded in the mental ability of the L1 learners such that their morphological understanding will bridge them in learning the target language. Finally, grammar of any language should be explored from the inside (Payne, 2021), and not from the outside perspective.

Appendices: The following are the linguistic symbols and abbreviations used in this study. They are presented in this paper alphabetically.

Appendix A. Symbols and Abbreviations

1	first person
2	second person
3	third person
s	singular
p	plural
-	affix boundary
=	clitic boundary
< >	infix
ABS	absolutive
ASP	Aspect
ERG	ergative
INTR	intransitive
LOC	locative
OBL	oblique
PAR	particle
PV	patient voice
TR	transitive

References

Cena, Resty (2014). A unified account of Tagalog verb and adjective affix system. In I. W. Arka, & M. Ni Luh Ketut , Argument realisation and related construction in Austronesian language (pp. 197-212). Australia: Asia- Pacific Linguistics.American Psychological Association. 1983. Publications Manual. American Psychological Association, Washington, DC.

Comrie, Bernard (1976). Aspect: An introduction to the study of the vebal aspects and related problems. Cambridge: Cambridge University Press.

Cunanan, Farrah C. (2015). Ang Dialect Area ng Bikol. Daluyan: Journal ng Wikang Filipino, 32-62.

Dery, Luis C. (1991). From Ibalon to Sorsogon: Historical Survey of Sorsogon Province to 1905. Quezon City: New Day Publisher.

Dita, Shirley N. (2010b). A morphosyntactic analysis of the pronominal system of the Philippine languages. 24th Pacific Asia Conference on Language, Information and Computation (pp. 45-59). Tokyo: PACLIC.

Escalante, Antonio (1978). A Study of South Sorsogon Verbs. Diliman, Quezon City:University of the Philippines.

Foley, William A. (1998). Symmetrical voice systems and precategoriality in Philippine languages. Paper presented at the 3rd LFG Conference, Brisbane.

Guzman, Videa P. (1994). Verbal affixes in Tagalog: Inflection or derivation. Seventh International Conference on Austronesian Linguistics (pp. 303-325). Leiden:Leiden Univ.

Haspelmath, Martin, & Sims, Andrea. D. (2010). Understanding Morphology: Understanding Language Series (second edition ed.). London : Hodder Education.

Himmelmann, Nikolaus P. (2008). Lexical categories and voice in Tagalog. In P. Austin, & S. Musgrave, Voice and Grammatical Functions in Austronesian Language (pp. 247-293). Stanford: CSLI.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online: http://www.ethnologue.com.

Lobel, Jason William, & Tria, Wilmer Joseph (2000). An Satuyang Tataramon: A study of the Bikol Language . Naga City: Lobel & Tria Partnership Co.

McFarland, Curtis Daniel (1974). The dialects of the Bikol area. Dissertation: Yale University.

Nolasco, Ricardo (2007). Si maka-tangkay at sim aka-ugat: Dalawang tagasuri ng morpolohiyang Pilipino. Paper presented during the 4th National Natural Language Processing Research Symposium, Philippine Languages and Computation on July 14-16, 2007. De La Salle University College of St. Benilde, Manila.

Nolasco, Ricardo D. (1994). Gramar ng Sorsoganon. Unpublished Masters' thesis. Diliman, Quezon City:University of the Philippines.

Payne, Thomas E. (1997). Describing Morphosyntax: A guide for field linguists . Cambridge: Cambridge University Press.

Payne, T. E. (2006). Exploring language structure: A student guide . New York: Cambridge University Press .

Payne, Thomas. E. (2021). Twin Course Webinar on L1 based MLE. SIL, Talaytayan, and BALLT Webinar via zoom.

Reid, L. A. (2016). Philippine languages- Pronominal System during the 2016 Linguistic Masterclass . National Museum. Manila: National Museum of Fine Arts.

Tanangkingsing, Michael (2009). A Functional Reference Grammar of Cebuano. Dissertation. National Taiwan University.

Zorc, David P. (1977). The Bisayan Dialect of the Philippines: Subgrouping and Reconstruction. Australia: Pacific Linguistics, s. C, no. 44.

Can the Translation Memory Principle Benefit Neural Machine Translation? A Series of Extensive Experiments with Input Sentence Annotation

Yaling Wang
Graduate School of IPS
Waseda University
Kitakyushu, Japan
yaling.wang@moegi.waseda.jp

Yves Lepage
Graduate School of IPS
Waseda University
Kitakyushu, Japan
yves.lepage@waseda.jp

Abstract

Integrating translation memories (TM) into neural machine translation (NMT) has been shown to improve translation quality. We test various schemes to integrate translation suggestions into an NMT system without altering its architecture. We retrieve similar sentences covering the sentence to translate and examine various annotation schemes as input to the NMT system. Our results show that the method can outperform a baseline model in some cases. The improvements are mainly for the translation of sentences with a length ranging from 10 to 20 words.

1 Introduction

Translation Memories (TMs) are used daily by translators. They contain aligned parallel sentence pairs. Given a sentence to translate, a TM retrieves the most similar sentence in the source language that contains large common or similar parts. The corresponding sentence in the target language is returned to the translator. In this way, the translator needs only to modify the unmatched parts to complete the translation. A main advantage of translation memories is that they ensure consistency and *interpretability* across translations because common or similar parts in sentences can easily be identified.

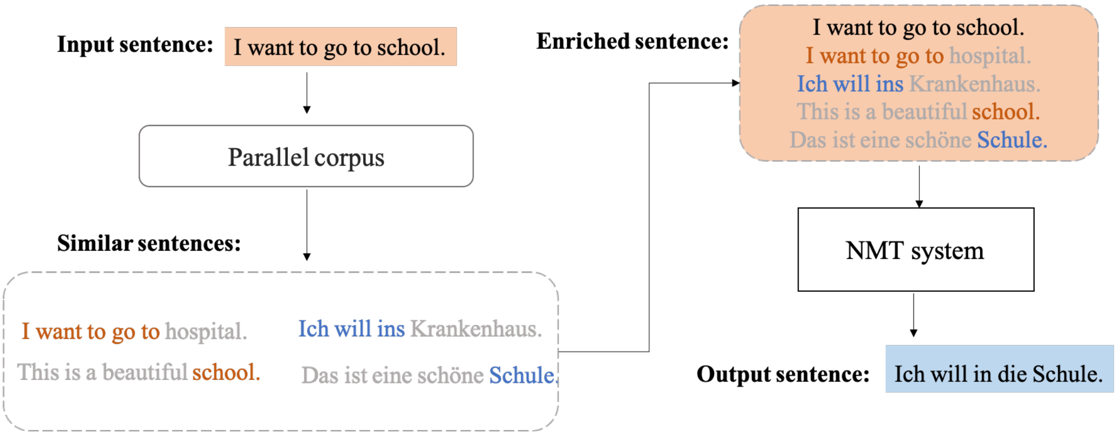
Recently, with the development of neural machine translation (NMT), the quality of machine translation has significantly increased. Its main advantage is that it improves translation *efficiency* over previous machine translation techniques. However, the interpretability of NMT is poor: errors are difficult to interpret, i.e., to trace back to the training data.

Past research (Federico et al., 2012) already proposed to combine the advantages of TM (interpretability) with MT (efficiency). Recently, methods have been proposed to achieve closer integration with NMT. For example, an additional encoder can be added to an NMT architecture specifically for TM matches (Cao and Xiong, 2018). The decoding algorithm can be modified to incorporate retrieved strings (Gu et al., 2018). An easy-to-implement TM–NMT integration has been proposed by (Bulté and Tezcan, 2019): they concatenate the target-language side of matches retrieved from a TM with the sentence to translate. This only involves data pre-processing and augmentation. This is also compatible with different NMT architectures. All of the approaches above were shown to lead to a significant increase in the quality of MT outputs.

2 Method

In this paper, we propose to make use of the TM principle in conjunction with an NMT system, without altering the model architecture of the NMT system. In this way, the method can apply to any neural network architecture and can leverage pre-trained models. Figure 1 illustrates this method.

Suppose that we want to translate a sentence from English to German. We firstly retrieve English sentences from the parallel aligned data and obtain some similar English sentences which cover the input English sentence. For instance, the sentence to translate '*I want to go to school.*' is covered by the two following sentences '***I want to go to** hospital.*'

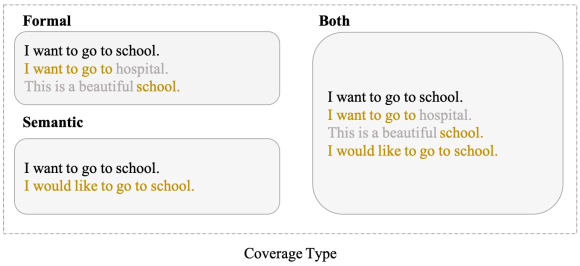


I want to go to hospital.
This is a beautiful school.
Ich will ins Krankenhaus.
Das ist eine schöne Schule.

I want to go to school.
I want to go to hospital.
Ich will ins Krankenhaus.
This is a beautiful school.
Das ist eine schöne Schule.

Figure 1: Overview of translation based on retrieval

and '*This is a beautiful **school.***'. Their corresponding German translations are obtained from the parallel data: '*Ich will ins Krankenhaus.*' and '*Das ist eine schöne Schule.*' That is, by retrieval, we acquire English–German sentence pairs, in which the English sentence is similar to the sentence to translate. The principle of translation memory postulates that the German sentences should also be similar to the German translation of the English sentence to translate. We use such translation pairs to enrich the input of an NMT system, i.e., we use them as annotations to the input sentence. We use such data to train an NMT system.



I want to go to school.
I want to go to hospital.
This is a beautiful school.

I want to go to school.
I would like to go to school.

I want to go to school.
I want to go to hospital.
This is a beautiful school.
I would like to go to school.

Figure 2: Illustration of different enrichment schemes that can be used according to the mode of retrieval (formal only, semantic only and both)

3 Enrichment Schemes

To emulate the principle of TM, we firstly retrieve a group of sentences in the source language that are similar to the sentence to translate. we use the tool introduced in (Liu and Lepage, 2021). Its goal is to cover a sentence in form and meaning with as few retrieved sentences as possible. It provides the possibility of retrieving sentences which are similar in form and meaning. Secondly, and to continue to emulate the TM principle, we obtain the corresponding sentences in the target part of the bilingual corpus.

However, we refine the principle of TM. The tool used for retrieval identifies the parts in the retrieved source sentences which are similar to the source sentence. Hence, based on these results, we use techniques in sub-sentential alignment from statistical machine translation, namely fast_align (Dyer et al., 2013), to obtain the corresponding translated parts in the sentences in the target language.

We now describe how we enrich the sentence to translate using results of retrieval and sub-sentential alignment. We propose different enrichment schemes to generate different possible inputs to the NMT system. Table 1 shows the different parameters which can be exploited under our settings. We describe them in details hereafter.

Coverage Type The tool used for retrieval provides two modes for similarity: formal and semantic similarity. Therefore, we can choose to retain sentences obtained by retrieval

- in form only;
- in meaning only;
- both.

Figure 2 illustrates results in these different modes.

Matched Parts Only or Whole Sentences From the retrieved sentences, we can choose to use:

Parameters	Options
Coverage Type	Formal only Semantic only Formal and Semantic
Matched parts	Matched parts only The whole sentences without markers The whole sentences with markers
Language Side	Source side only Target side only Source and Target sides
Order of Similar Sentences	All source sentences followed by all target sentences All target sentences followed by all source sentences Each source sentence followed by its corresponding target sentence, for all pairs of sentences Each target sentence followed by the source sentence it corresponds to, for all pairs of sentences

Table 1: List of different parameters that can be exploited to produce different enrichment schemes

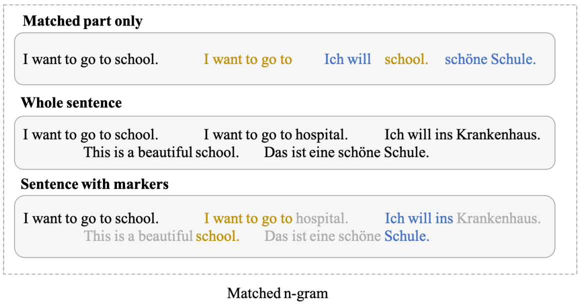

Matched part only
I want to go to school.
I want to go to
Ich will
school.
schöne Schule.
Whole sentence
I want to go to school.
I want to go to hospital.
Ich will ins Krankenhaus.
This is a beautiful school.
Das ist eine schöne Schule.
Sentence with markers
I want to go to school.
I want to go to hospital.
Ich will ins Krankenhaus.
This is a beautiful school.
Das ist eine schöne Schule.
Matched n-gram

Figure 3: Illustration of different enrichment schemes that can be used: with matched parts or whole sentences, with markers or not

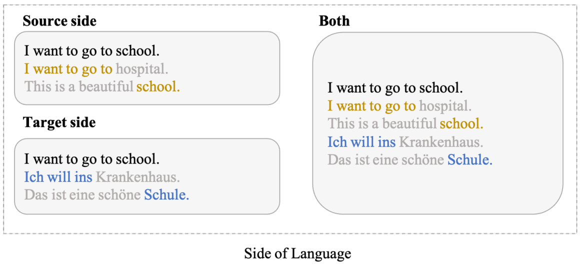

Source side
I want to go to school.
I want to go to hospital.
This is a beautiful school.
Target side
I want to go to school.
Ich will ins Krankenhaus.
Das ist eine schöne Schule.
Both
I want to go to school.
I want to go to hospital.
This is a beautiful school.
Ich will ins Krankenhaus.
Das ist eine schöne Schule.
Side of Language

Figure 4: Illustration of different enrichment schemes that can be used depending on the language side used: source side only, target side only and both)

- only the matched parts, i.e., the parts which are similar to the sentence to translate (in form or in meaning, directly in the source language or by translation and sub-sentential alignment in the target language), and only these parts;
- the whole sentences retrieved with markers so as to identify the matched parts, in the source or the target language;
- the whole sentences retrieved without any markers to identify the matched parts.

Figure 3 provides an illustration of such possible cases.

Language Side Following the principle of TM, we obtain similar sentences in the source language by retrieval. Now, the corresponding sentences in the target language should contribute to translation. In terms of language side, we can thus choose

- the sentences retrieved in the source language only;
- the corresponding translations in the target language only;
- both: the sentences in the source language and their corresponding translations in the target language.

Figure 4 illustrates the above three possibilities.

Order of Similar Sentences When both language sides are chosen, we can imagine several enrichment schemes for the ordering of the sentences in the source language and the target language.

- All source sentences followed by all target sentences;
- All target sentences followed by all source sentences;
- Each source sentence followed by its corre-

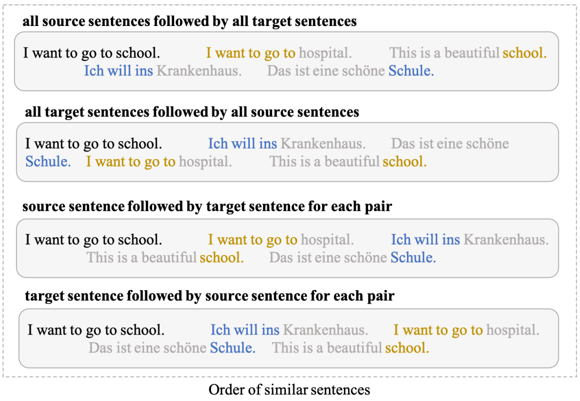

all source sentences followed by all target sentences
I want to go to school. I want to go to hospital. This is a beautiful school. Ich will ins Krankenhaus. Das ist eine schöne Schule.
all target sentences followed by all source sentences
I want to go to school. Ich will ins Krankenhaus. Das ist eine schöne Schule. I want to go to hospital. This is a beautiful school.
source sentence followed by target sentence for each pair
I want to go to school. I want to go to hospital. Ich will ins Krankenhaus. This is a beautiful school. Das ist eine schöne Schule.
target sentence followed by source sentence for each pair
I want to go to school. Ich will ins Krankenhaus. I want to go to hospital. Das ist eine schöne Schule. This is a beautiful school.
Order of similar sentences

Figure 5: Illustration of different enrichment schemes that can be used for the ordering of similar sentences when both language sides are used

sponding target sentence for all pairs of sentences;

- Each target sentence followed by the source sentence it corresponds to, for all pairs of sentences.

Figure 5 shows an example.

List of all Possible Enrichment Schemes All possible choices for each of the parameters enumerated above lead to a list of 54 possible enrichment schemes for the exploitation of the information obtained following the TM principle. Table 5 lists them all.

4 Experimental Setup

4.1 Data

We use Multi30k (Elliott et al., 2016) as our parallel corpus. It contains multilingual image descriptions for multilingual and multimodal research. We use the German, English and French parts in our experiments. Some statistics are given in Table 2. We split the dataset into 3 parts, 80% for training, 10% for validation and 10% for testing. We perform translation experiments in all possible directions offered by the three languages, i.e., 6.

4.2 Evaluation

Following standard practice, evaluation of translation is done by computing the BLEU score (Papineni et al., 2002) on the test set. We report BLEU scores in the range of 0 to 100. BLEU scores indicate similarity to the reference translation in form only.

Hence, in addition to BLEU scores, for the purpose of measuring semantic similarity with the reference, we compute BERTScores (Zhang et al., 2020). BERTScores leverage pre-trained contextual embeddings from BERT and match words in the candidate and the reference sentences using cosine similarity. We report the F-measure, which ranges from 0 to 1. Higher BERTScores indicate higher similarity in meaning between the candidate and the reference sentences.

lang.	# sents.	vocab. size	avg. length (in tokens)
de	30,014	18,722	12.44
en	30,014	10,214	13.02
fr	30,014	11,794	13.62

Table 2: Statistics on the corpus (Multi30k)

Encoder	
Type	LSTM
Embedding Dimension	500
Number of layers	2
Size of hidden layer	500
Decoder	
Type	StackedLSTM
Embedding Dimension	500
Number of layers	2
Size of hidden layer	500
Total # of parameters	18,368,003
Optimizer	SGD
Learning rate	1.0

Table 3: Configuration for the NMT model

4.3 Baseline System

We compare our proposal to a baseline. Our baseline model is trained using the same NMT model but simply with the sentences to translate without anything else, as input.

Our NMT model follows the Seq2seq architecture (Bahdanau et al., 2015) implemented in the OpenNMT-py toolkit (Klein et al., 2017). The model configuration is shown in Table 3.

5 Experiment Results

5.1 Results for Retrieval

For each of the sentences in the English, German and French test sets, we apply the TM principle and retrieve similar sentences.

Retrieval in Form For the results of retrieval in form, we focus on the number of similar sentences retrieved per input sentence. This is because the retrieval method used aims at maximal coverage with the least possible number of retrieved sentences. A lesser number of similar sentences means that the common parts are longer.

Table 4 gives some statistics on the results of retrieval. The number of retrieved sentences in the 3 languages is similar. The average number of retrieved sentences is about 5, which means that 5 n-grams in the retrieved sentences almost cover the input sentence. The value of the standard deviation is also relatively small. The most frequent number of retrieved sentences is 4, 5 or 6. There are only few cases where the number of retrieved sentences is greater than 10. This means that, in general, the retrieval method used covers the sentence to translate with a relatively small number of similar sentences.

Retrieval in Meaning Table 4 gives some statistics on the results of semantic retrieval. Compared to retrieval in form, the number of retrieved sentences is less. This is because the method used selects the top k sentences that contribute to the increase in coverage of the input sentence. These sentences are a supplement.

5.2 Translation Results with Different Enrichment Schemes

We measure the performance of different enrichment schemes and select the scheme that performs the best. The translation task that we consider is from German to English, so that we use the German sentences for the retrieval step. Figure 8 shows examples of translations obtained using different enrichment schemes and Table 5 gives the results of evaluation for all possible different enrichment schemes.

To analyze the results, we draw box plots by groups of four parameters (coverage type, matched parts, side and ordering), in Figures 6 and 7. The box plot show six ranges for the results: upper edge, upper quartile, median, lower quartile, lower edge, and outliers.

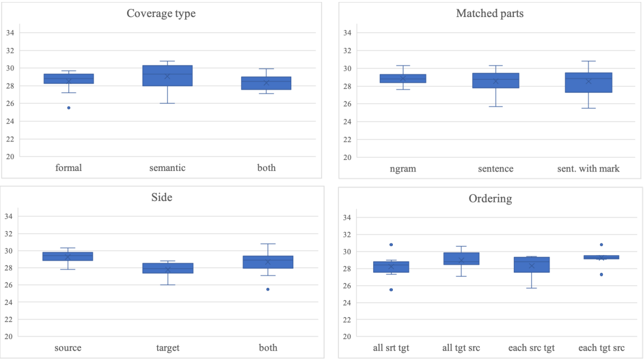


Figure 6: Box plots by coverage type, matched part/whole sentence, language side and ordering on BLEU scores.

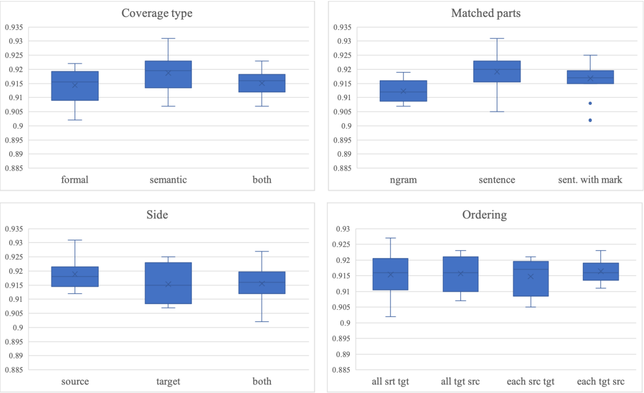


Figure 7: Box plots by coverage type, matched part/whole sentence, language side and ordering on BERTScores.

In terms of coverage type, among three coverage types, the median BLEU scores of formal coverage and semantic are similar. The BERTScore with semantic coverage is the highest. This is expected because for semantic retrieval, retrieves sentences that have similar meanings by definition.

The results of only using matched parts for translation is found to be the most stable with the smallest standard deviations. However, the performance across the three possible choices (only matched part, sentence, sentence with marker) is close in scores.

In terms of language side, among the three possible options, using source information only performs the best in BLEU with an average score of 29.26.

Retrieval	Language	# of retrieved sentences			Average length	
		mean ± stdev.	median	mode	in tokens	in char.
Formal	de	5.03 ± 2.22	5	4	62	352
	en	5.50 ± 2.22	5	5	71	332
	fr	5.40 ± 2.32	5	4	76	395
Semantic	de	2.81 ± 1.42	3	2	36	209
	en	2.22 ± 1.14	2	2	29	137
	fr	2.69 ± 1.33	2	2	38	199

Table 4: Statistics of results for formal (top) and semantic retrieval (bottom)

No.	Retrieval	Matched parts	Side	Ordering	BLEU	BERTScore
1	Formal only	n-gram	Source only	-	28.8	0.923
2		sentence		-	29.5	0.926
3		sent. with markers		-	29.7	0.927
4		n-gram	Target only	-	28.3	0.919
5		sentence		-	28.7	0.920
6		sent. with markers		-	27.2	0.917
7		n-gram	Source and Target	all src. tgt.	28.2	0.918
8		n-gram		all tgt. src.	28.4	0.921
9		n-gram		each src. tgt.	29.3	0.925
10		n-gram		each tgt. src.	29.1	0.923
11		sentence		all src. tgt.	28.5	0.927
12		sentence		all tgt. src.	29.4	0.926
13		sentence		each src. tgt.	25.7	0.914
14		sentence		each tgt. src.	29.3	0.926
15		sent. with markers		all src. tgt.	25.2	0.912
16		sent. with markers		all tgt. src.	28.8	0.925
17		sent. with markers		each src. tgt.	29.4	0.925
18		sent. with markers		each tgt. src.	29.3	0.923
19	Semantic only	n-gram	Source only	-	30.3	0.923
20		sentence		-	29.4	0.932
21		sent. with markers		-	28.9	0.926
22		n-gram	Target only	-	27.9	0.924
23		sentence		-	28.0	0.927
24		sent. with markers		-	26.0	0.927
25		n-gram	Source and Target	all src. tgt.	29.0	0.921
26		n-gram		all tgt. src.	28.5	0.918
27		n-gram		each src. tgt.	29.3	0.920
28		n-gram		each tgt. src.	29.3	0.921
29		sentence		all src. tgt.	27.8	0.929
30		sentence		all tgt. src.	30.3	0.928
31		sentence		each src. tgt.	27.6	0.926
32		sentence		each tgt. src.	29.5	0.927
33		sent. with markers		all src. tgt.	30.8	0.925
34		sent. with markers		all tgt. src.	30.6	0.928
35		sent. with markers		each src. tgt.	29.4	0.924
36		sent. with markers		each tgt. src.	30.8	0.925

No.	Retrieval	Matched parts	Side	Ordering	BLEU	BERTScore
37	Formal and Semantic	n-gram	Source only	-	29.9	0.926
38		sentence		-	27.8	0.921
39		sent. with markers		-	29.0	0.925
40		n-gram	Target only	-	27.6	0.917
41		sentence		-	28.8	0.927
42		sent. with markers		-	27.6	0.923
43		n-gram	Source and Target	all src. tgt.	28.6	0.923
44		n-gram		all tgt. src.	28.6	0.918
45		n-gram		each src. tgt.	28.8	0.919
46		n-gram		each tgt. src.	29.2	0.921
47		sentence		all src. tgt.	27.3	0.921
48		sentence		all tgt. src.	29.0	0.924
49		sentence		each src. tgt.	27.9	0.925
50		sentence		each tgt. src.	29.5	0.925
51		sent. with markers		all src. tgt.	28.4	0.925
52		sent. with markers		all tgt. src.	27.1	0.923
53		sent. with markers		each src. tgt.	27.5	0.923
54		sent. with markers		each tgt. src.	27.3	0.924

Table 5: All possibilities of formats with results of evaluation. All confidence intervals for the BLEU scores are between 0.75 and 0.85.

As for the order of similar sentences, the second order (all target sentences followed by all source sentences) and the fourth order (each target sentence followed by the source sentence it corresponds to, for all pairs of sentences) perform better than the other ones in BLEU. This shows that giving target information before source information is a better choice.

All in all, to select the best combination of four parameters among all the possible formats through BLEU score and BERTScore, we notice that a higher BLEU score is not always accompanied by a higher BERTScore. We want the translations to be close to the reference translations not only in form but also in meaning. Hence, we sort all configurations using the average of the BLEU scores (recast from 0 to 1) and BERTScores and select the configuration ranked the highest. It is configuration No. 36. We apply this best configuration in all other translation directions.

5.3 Translations in Different Languages

We perform machine translation experiments in all directions of all languages pairs between German, English and French. This is 6 directions in total. We use enrichment scheme No. 36, i.e., results of semantic retrieval only, using whole sentences with matching parts indicated with markers, each target sentence followed immediately by the source sentence it corresponds to, for all pairs of retrieved sentences. Table 6 summarizes the translation results.

When using formal coverage retrieval results, our models outperform the baseline model in three translation tasks: de→en, de→fr and fr→en. In the other cases, although our models do not exceed the baseline system, confidence intervals, as shown in Table 6, indicate that the baseline model and our models perform similarly. For instance, for the direction en→de, confidence intervals of ± 0.8 do not allow to say that a baseline of 27.4 is really better than our model with 27.1. As the main difference is the language of query sentences, i.e., the source language, we might think that the differences in BLEU observed by the difference in morphology of the source and target languages explain the results. In general, the result shows that the formal coverage retrieval method contributes to improving the translation quality or performs similarly compared to the baseline system.

When using semantic coverage retrieval, our

Translation direction			Baseline	Proposed method	
				Formal coverage	Semantic coverage
de	→	en	29.6 ± 0.8	**30.5** ± 0.9	**30.6** ± 0.8
de	→	fr	30.6 ± 0.8	**31.8** ± 0.8	29.7 ± 0.8
en	→	de	27.4 ± 0.8	27.1 ± 0.8	26.1 ± 1.0
en	→	fr	42.2 ± 1.2	41.8 ± 1.2	**47.2** ± 1.0
fr	→	de	24.3 ± 0.8	24.1 ± 0.8	23.6 ± 0.8
fr	→	en	38.8 ± 0.9	**39.6** ± 0.9	**42.5** ± 1.2

Table 6: Translation results (in BLEU) for each different translation directions

No.	Sentence to translate	Output translations	Reference translation
3	ein mann in einem gelben oberteil macht eine inspektion an einem schwinn-fahrrad neben einem picknicktisch .	a man in a yellow top is making a inspektion at a schwinn-fahrrad .	man in yellow shirt performing maintenance on schwinn bicycle near a picnic table .
9		a man in a yellow top is taking a break by a picnic table next to a picnic table .	
15		a man in a yellow top is taking a trick by a picnic table .	
41	ein mann auf einem motorrad und zwei weitere männer auf einem wagen fahren auf einer staubigen zweispurigen straße .	a man on a motorcycle and two other men on a wagen on a sunny road .	a very man on a motorcycle and 2 men on a cart are traveling down a dusty two lane road .
35		a man on a motorcycle and two other men riding on a dusty bike .	
38		man on a motorcycle and two more men on a dusty road .	
42	ein ball befindet sich zwischen einem werfer und einem fänger auf dem baseballfeld .	a ball is in between a werfer and batter on the baseball field .	a pitcher and catcher on a baseball field with the ball in between them .
41		a ball is between a werfer and baseball on the baseball .	
40		a ball is between a werfer and a fänger on the baseball .	

Figure 8: Examples of translation using different formats

Input sentence	**Translation**	**Reference**
one lady in a plaid coat eating cotton candy .	une femme en manteau à carreaux mange de la barbe .	une femme en veste écossaise mangeant de la barbe à papa .
two men and a woman are inspecting the front tire of a bicycle .	deux hommes et une femme untersuchen le vorderrad d&apos; un vélo .	deux hommes et une femme inspectent le pneu avant d&apos; un vélo .
un petit chien avec un ruban rouge sur sa tête marche dans l&apos; herbe .	ein kleiner hund mit einer roten ruban auf seinem kopf .	ein kleiner hund mit einem roten band auf dem kopf läuft durch das gras .
trois femmes en rouge de l&apos; équipe de basket russe suivant le ballon .	drei frauen in roter équipe suivant suivant .	drei frauen in roten trikots aus der russischen basketball-mannschaft laufen dem basketball hinterher .
ein thaiboxer übt zum aufwärmen vor dem kampf einen beinhochtritt .	a thaiboxer band is practicing for the aufwärmen in front of the net .	this thai boxer is practicing a high leg kick as a warm up before his fight .
ein mann mit einem rucksack springt von einem pier .	a man with a backpack jumps off a pier .	a man wearing a backpack is jumping off a pier .

Figure 9: Random examples in different translation directions

models outperform the baseline model in three translation tasks: de→en, en→fr and fr→en. This is the same number as for formal coverage, but one language direction is different: en→fr instead of de→fr. A large improvement is obtained in the direction: fr→en. In this translation task, the model using semantic coverage retrieval outperforms the baseline model by 3.7 BLEU points, which is largely more than the model using formal coverage retrieval. Our method leads to an even larger improvement in the translation task en→fr using semantic coverage retrieval. The BLEU score increases by 5.0 points over the baseline model, whereas the model using formal coverage retrieval does not exceed the baseline system. We conclude that our proposed method with semantic coverage is especially efficient for the language pair en–fr, in both directions.

Figure 9 shows some examples of translation results. (input sentence is just source sentence without enrichment)

5.4 Length of the sentence to translate

Based on some samples, we found that our model delivers similar performance as the baseline model for shorter sentences (length less than ten words). However, our model offers better translations for sentences between 10 and 20 words due to the information found in similar sentence pairs. In order to confirm the impression left by this observation, we split the test set into three parts by the length of the sentence to translate, and we compare the performance on these three separate subsets.

Length of sentences	# of sentences	BLEU score	
		Baseline	Ours
<10	448	31.3	31.0
10–20	2,207	30.1	**31.1**
>20	247	25.9	25.5

Table 7: Translation results for different sentence lengths (in BLEU, de→en)

Table 7 shows the results for the three separate subsets containing sentences with different lengths. The sentences of a length between 10 and 20 words account for the most part of the test set. Our model outperforms the baseline model on this subset by 1.0 BLEU point. However, for sentences of length more than 20, both models cannot perform well.

6 Conclusion

We proposed to test whether the principle of translation memory (TM) can benefit results in neural machine translation (NMT). We enriched the input of the NMT system with such sentences retrieved. We studied different annotation schemes, and found that the scheme which delivers the best translation accuracy consists in providing the target sentence immediately before its corresponding source sentence, for all sentence pairs, and identifying matching parts with markers.

Such enrichment schemes can contribute to the interpretability of the results obtained by neural machine translation systems. The results of our translation experiments show that, for some translation tasks, our system performs better than a standard NMT system without retrieval. Increases in translation accuracy are mainly obtained for sentences with a length in the range of 10 to 20 words.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bram Bulté and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, Brussels, Belgium. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search engine guided neural machine translation. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pages 5133–5140. AAAI press.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Yuan Liu and Yves Lepage. 2021. Covering a sentence in form and meaning with fewer retrieved sentences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 35)*, pages 1–10.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

To Post or Not to Post: Exploring Students' Writing Strategies in a COVID-stricken Classroom

Jennifier T. Diamante
Western Philippines University
Palawan, Philippines
jennifier.diamante@wpu.edu.ph

Romualdo A. Mabuan
Far Eastern University
Manila, Philippines
rmabuan@feu.edu.ph

Emmalyn T. Venturillo
Western Philippines University
Palawan, Philippines
mmventurillo@wpu.edu.ph

Abstract

This study explores the affordances of microblogging as a strategy in improving the writing proficiency of students. A total of 125 college freshman students from a state university in the Philippines served as the respondents of the study. Four writing prompts were used for microblogging and personal essay writing. UAM Corpus Tool was used to describe the linguistic characteristics of the writing compositions produced by the respondents, while paired t-test and Mann-Whitney U test were used to statistically analyze the data. A Discourse Completion Task (DCT) was also employed to generate qualitative information to deepen the interpretation of the findings. To validate the ratings given by the researcher, an inter-rater who is also an English professor was invited to rate the data. Holistically, there is no statistical difference between the mean ratings obtained by the respondents in the two writing conditions. Analytically, there is a highly significant difference in sentence structure and a significant difference in terms of language use. Moreover, micro-blogging and personal writing have shown no significant difference in content, organization, and mechanics as revealed by the statistical results. Conclusions and recommendations were forwarded. Pedagogical implications were articulated based on the results of the investigation.

1 Introduction

The COVID-19 pandemic has disrupted not only the economic, political, and health-care system but also the academic set-up around the globe. The lockdown in response to the virus has forced educational institutions worldwide to adopt learning modalities that they deemed most appropriate and realistic for their learners given their available resources. As the new mode of teaching seems challenging to the teachers, its benefits to students are also being questioned. On the one hand, teachers had to adapt to the new pedagogical concepts and teaching mode delivery for which they may not have been trained. On the other hand, students, especially those in the marginalized groups, who do not have access to the digital learning resources or lack of resilience and engagement to learn on their own, are at risk of falling behind (Schleicher, 2020).

In the Philippines, higher educational institutions (HEIs) had to replace face-to-face lectures with online and modular learning along with other modalities which they deemed necessary and appropriate for their contexts to ensure the safety of both teachers and students. While shifting to the new normal of teaching and learning poses great challenges to various stakeholders especially to the teachers and students, it creates an opportunity to explore new possibilities in pedagogy and research. This study, therefore, investigates how students perform academically in the digitalized learning environment particularly in their writing assignments.

In the context of English as Second Language (ESL) teaching, writing is the most difficult English skill (Richards & Renandya, 2002) and the most difficult communicative macro skill to teach and master (Mabuan, 2018; Tillema, 2012). Perhaps, it has become more challenging using an unexplored web platforms.

Google Classroom (GCR) is one of the platforms that is being used by WPU for its online learning, but its functionality has not yet to be explored from the lens of language research. GCR is a website which also enables teachers to create and manage the online classroom as if it were a conventional one where they can share learning materials, post assignments, prepare quizzes, and give instructions to their learners. Students, on the other hand, can access these materials and complete the task as long as they have Internet connectivity. Thus, this study takes advantage of the possibility that students' writing outputs published in the Stream Section of the GCR, coined in this paper as microblogging, is an effective strategy that could have impact on students' writing skill. This research study problematizes whether microblogging and personalized essay writing where students write and submit their compositions directly to their teachers has significant impact on the quality of students' output.

1.1 Research Questions

This study sought to address the following questions:

1. What are the lexical features of students' essays?
2. What is the impact of micro-blogging and personalized essay writing to students' writing ability?
3. What are the affordances of micro-blogging to students' writing ability?

2 Method

2.1 Participants

This study involved four sections of 125 freshman students enrolled in the Bachelor of Science and Social Work (BSSW) program of Western Philippines University-Puerto Princesa Campus in Palawan, the Philippines during the first semester of school-year 2020-2021 in the Purposive Communication Course, one of the general education courses under the new curriculum that is being handled by the researcher.

2.2 Procedure

2.2.1 Topic Validation

For topic validation, the four writing prompts used in the study were presented to a colleague who is also an English faculty and to another English instructor from other institution who also served as an interrater of the data.

2.2.2 Administration of the Writing Tasks

Under the COVID-19 pandemic, instructional modules were developed by the researcher and her colleagues as the main resource uploaded to the Google Classroom bi-monthly. For each module, the researcher included a writing task which is related to a lesson discussed in the learning packet that will be submitted by the students within the two-week period specified in their course outline. The first writing prompt follows (Please refer to Appendix A for the complete list).

2.2.2.1 For microblogging

Write a minimum of 500 and maximum of 1000 word-essay on the topic "Are you the same person of social media as you are in real life?" Make sure to provide appropriate details in the introduction, body, and conclusion sections of your essay. After you submit your essay to the file I created for this assignment, post it to the stream section of our Google Classroom for the whole class to read your work. Please make sure to also provide feedback on the work of your classmates.

2.2.2.2 For traditional essay submission

Write a minimum of 500 and maximum of 1000 word-essay on the topic "Are you the same person of social media as you are in real life?" Make sure to provide appropriate details in the introduction, body, and conclusion sections of your essay. Submit your essay to the file I created for this assignment.

All essays were holistically rated by the researcher and the inter-rater using Gustilo's (2011) rubric. Gustilo used the said rubric in a study where she assessed 150 essays from five universities in the Philippines, and in another that examines the writing performance of engineering students in a prestigious tertiary institution in the Philippines (Gustilo, 2011, 2013). Only those essays that were submitted on time and have negligible traces of plagiarism were analyzed, the rest were discarded.

2.2.3 Reflective Survey

After completing the four writing tasks, the students were given an opportunity to reflect on whether or not microblogging affects the quality of their output. They were asked to accomplish a Discourse Completion Task (DCT) requiring them to answer reflective questions (Please see Appendix D for some DCT student responses).

2.2.4 Analysis

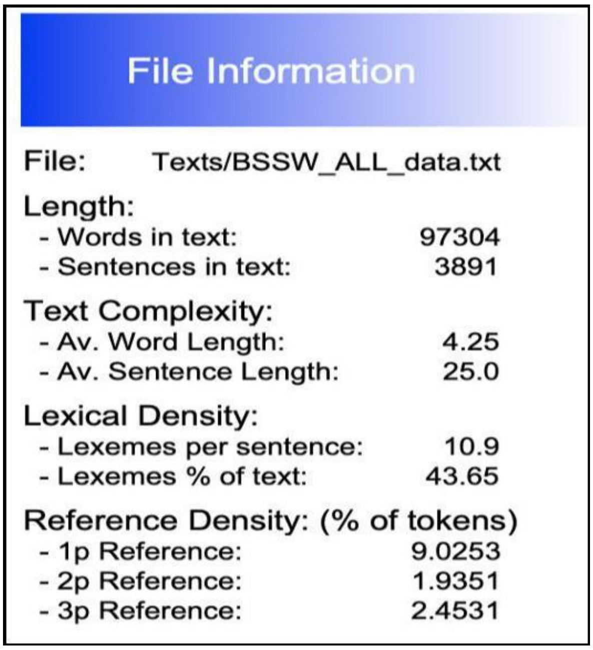
1. The UAM Corpus Tool, a software used for linguistic tagging was used to provide the characteristics of the data.
2. Surveys were conducted via Google forms hence replies were automatically tallied in terms of frequencies.
3. Essay scores from the two raters were averaged if there is a one-point discrepancy.
4. Qualitative replies were analyzed and used as reference whenever necessary to deepen the interpretation of the quantitative data.

3 Results and Discussion

3.1 Lexical Features

Figure 1 presents the characteristics of students' essays in terms of length, text complexity, lexical and reference densities.

Figure 1
Data Characteristics



```
File Information

File:      Texts/BSSW_ALL_data.txt
Length:
  - Words in text:              97304
  - Sentences in text:           3891
Text Complexity:
  - Av. Word Length:             4.25
  - Av. Sentence Length:         25.0
Lexical Density:
  - Lexemes per sentence:        10.9
  - Lexemes % of text:          43.65
Reference Density: (% of tokens)
  - 1p Reference:              9.0253
  - 2p Reference:              1.9351
  - 3p Reference:              2.4531
```

The figure shows that the data is consists of 3,891 sentences with a total of 97,304 word-tokens. As can be seen, the data has the average the word length of 4.25 letters and the average sentence length of 25 words. Figure 1 likewise shows that the average lexemes per sentence is 10.9, while the entire text contains 43.65 per cent of lexemes. In terms of reference density, first-person pronouns were generally used by the respondents, which is given since two of the writings prompts were about the students' reflections and experiences amidst the COVID-19 pandemic.

What is compelling in the findings is the tendency of the students to write longer sentences with somewhat shorter words. As shown in Figure 1, the average sentence length of the data is 25 words and the average word length of 4.25 letters. In (1), the sentence has 25 words, (2) has 34 and
(3) contains 50 words.

(1) Hey it's been months since the pandemic started, a pandemic that affect not only the people in cities and provinces but also in our island.

(2) To: My Family & friends I hope at this time You are okay and your health is good and far

from disaster especially with the ongoing disaster that is spreading all over the world.

(3) My personality is not the same as what you can see on my social media, because for me social media is just a consolation where I can share posts even if it is not really about me, it is very different from my real life as it reflects my personality.

3.2 Students' Writing Proficiency

The writing proficiency of the respondents is summarized in Table 1.

Table 1
Writing Proficiency of the Respondents

Sections	Micro-blogging (Posted) Mean	Adjectival Rating	Traditional (Not posted) Mean	Adjectival Rating
A	3.87	Developing Proficiency	4.06	Adequate Proficiency
B	4.12	Adequate Proficiency	4.1	Adequate Proficiency
C	4.29	Adequate Proficiency	4.29	Adequate Proficiency
D	4.3	Adequate Proficiency	4.20	Adequate Proficiency
Grand Mean	**4.05**	**Adequate Proficiency**	**4.16**	**Adequate Proficiency**

Legend: 1 = Very Little Proficiency 2 = Little Proficiency 3= Developing Proficiency
4= Adequate Proficiency 5 = Advanced Proficiency 6 = Highly Advanced Proficiency

The table shows that the writing proficiency of the respondents both for the microblogging and traditional essay writing can be generally described as having adequate proficiency. In the TOEFL Test of English Scoring Guide (ETS, 2019), an essay with a rating of 4 is adequately organized, addresses the writing topic adequately but may slight parts of the task, uses some details to support a thesis or illustrate ideas, uses adequate but undistinguished or inconsistent facility with syntax and usage may contain some serious errors that occasionally obscure meaning. This means that the writer in this category demonstrates minimal competence on both the rhetorical and syntactic levels.

3.3 Microblogging versus Traditional Writing

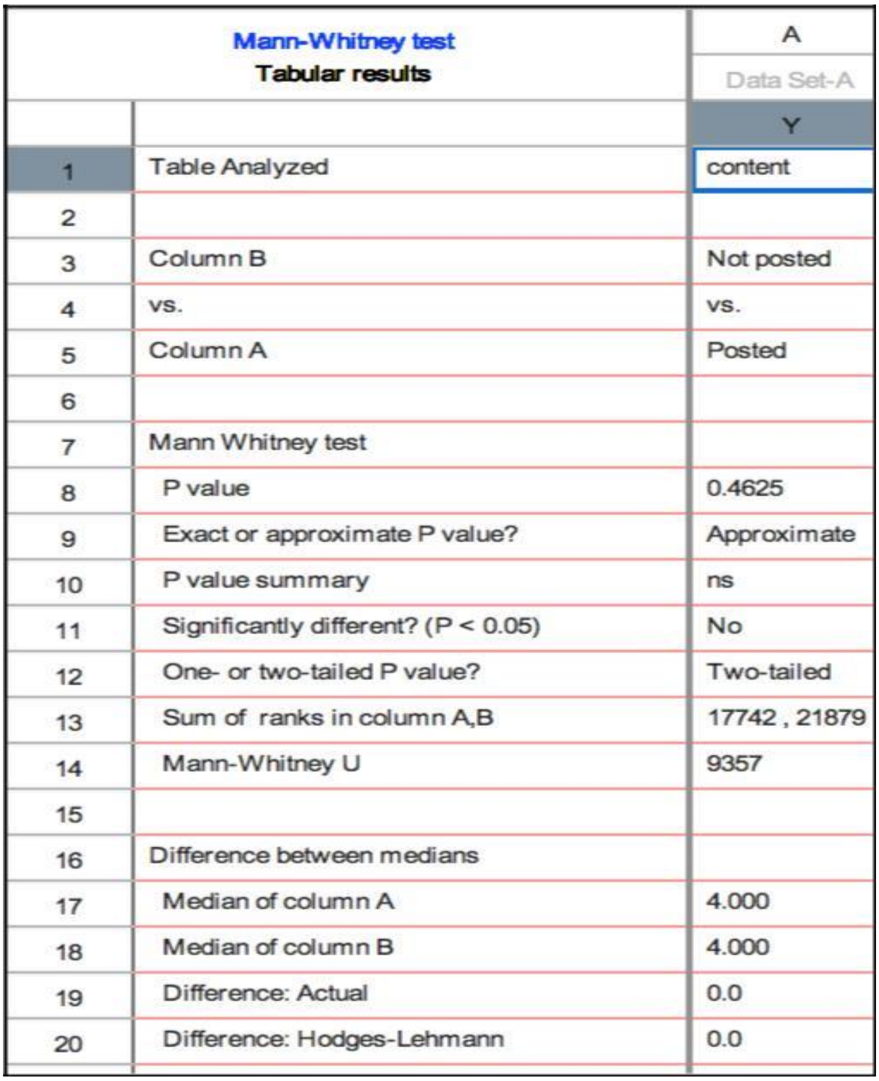
Answers to the question on whether microblogging and traditional essay writing employed as strategies of the teacher in the writing classes are presented in Tables 2 to Table 6.

3.4 On Content

Content refers to sound information, adequate and appropriate details provided by the respondents in their essays. Table 2 presents that there is no significant difference between microblogging and the traditional essay writing in terms of content. This finding is expected since the respondents whether they micro-blog or submitted their essay to their teacher wrote about the same topics. Given that the respondents share a similar demographic profiles, it is more likely that they also have similar concepts and ideas on the topics assigned to them by their teacher.

Table 2

Mann-Whitney Tabular Results on Content



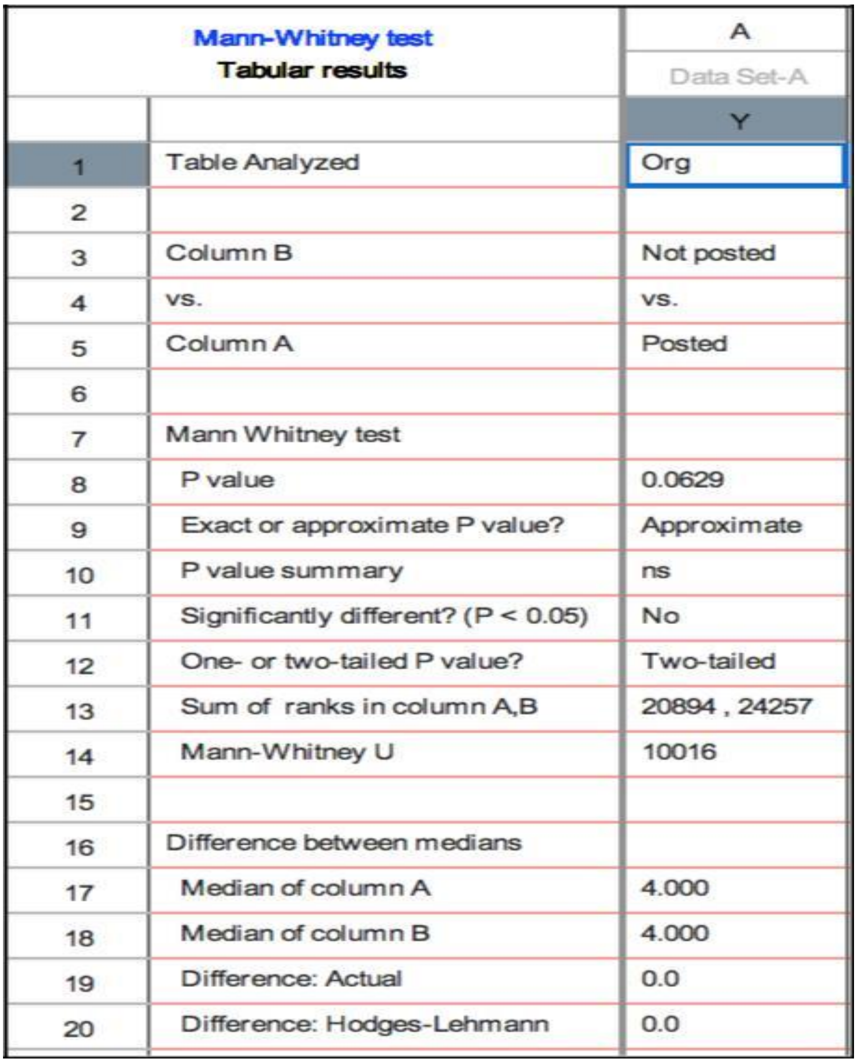
Mann-Whitney test Tabular results		A Data Set-A
		Y
1	Table Analyzed	content
2		
3	Column B	Not posted
4	vs.	vs.
5	Column A	Posted
6		
7	Mann Whitney test	
8	P value	0.4625
9	Exact or approximate P value?	Approximate
10	P value summary	ns
11	Significantly different? (P < 0.05)	No
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	17742 , 21879
14	Mann-Whitney U	9357
15		
16	Difference between medians	
17	Median of column A	4.000
18	Median of column B	4.000
19	Difference: Actual	0.0
20	Difference: Hodges-Lehmann	0.0

3.5 On Organization

Respondents are expected to have skillfully arranged their ideas in the introduction, body, and conclusion parts of their essay. The Mann-Whitney results on organization shows no significant difference between the essays produced in microblogging and in traditional essay writing as can be gleaned from Table 3. According to Nordquist (2020), organization in speech or composition is the arrangement of ideas, incidents, evidence, or details in a perceptible order in a paragraph, essay or speech. In classical rhetoric, organization is also known as the elements' arrangement or *dispositio* (Nordquist, 2020).

Table 3

Mann-Whitney Tabular Results on Organization



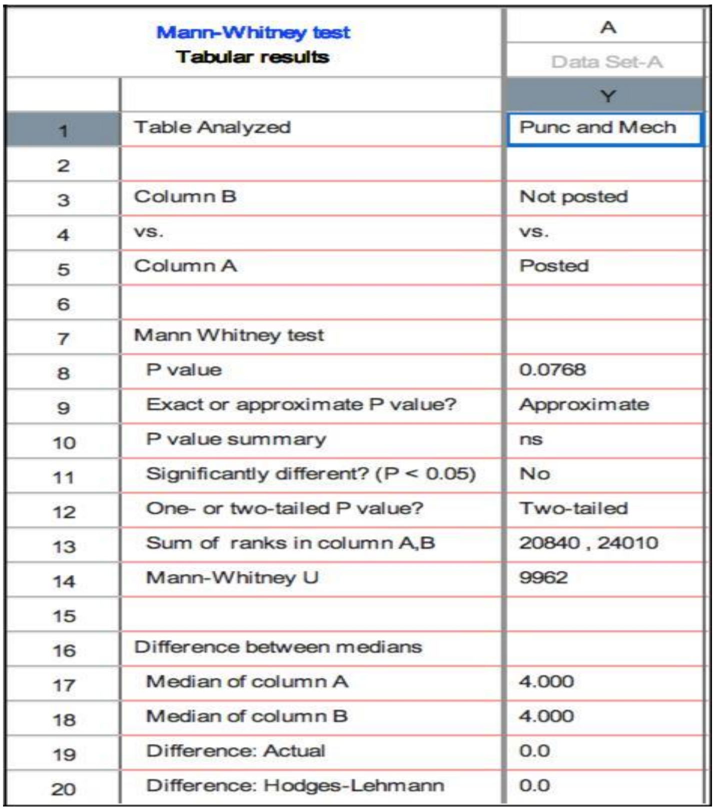
Mann-Whitney test Tabular results		A Data Set-A
		Y
1	Table Analyzed	Org
2		
3	Column B	Not posted
4	vs.	vs.
5	Column A	Posted
6		
7	Mann Whitney test	
8	P value	0.0629
9	Exact or approximate P value?	Approximate
10	P value summary	ns
11	Significantly different? (P < 0.05)	No
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	20894 , 24257
14	Mann-Whitney U	10016
15		
16	Difference between medians	
17	Median of column A	4.000
18	Median of column B	4.000
19	Difference: Actual	0.0
20	Difference: Hodges-Lehmann	0.0

3.6 On Punctuation and Mechanics

Like content and organization, usage of punctuations and mechanics show no significant difference between the compositions written under the two writing conditions. Two interpretations could be deduced from the findings. One, the respondents failed to master the rules punctuation usage along with capitalizations, indentions, spelling, and grammar. Two, some of these errors manifested in the compositions are not due to lack of knowledge of rules but is simply caused by carelessness thus resorted to mistakes but not necessarily errors, as echoed in the study of Mohammadi and Mustafa (2020).

Table 4

Mann-Whitney Tabular Results on Mechanics

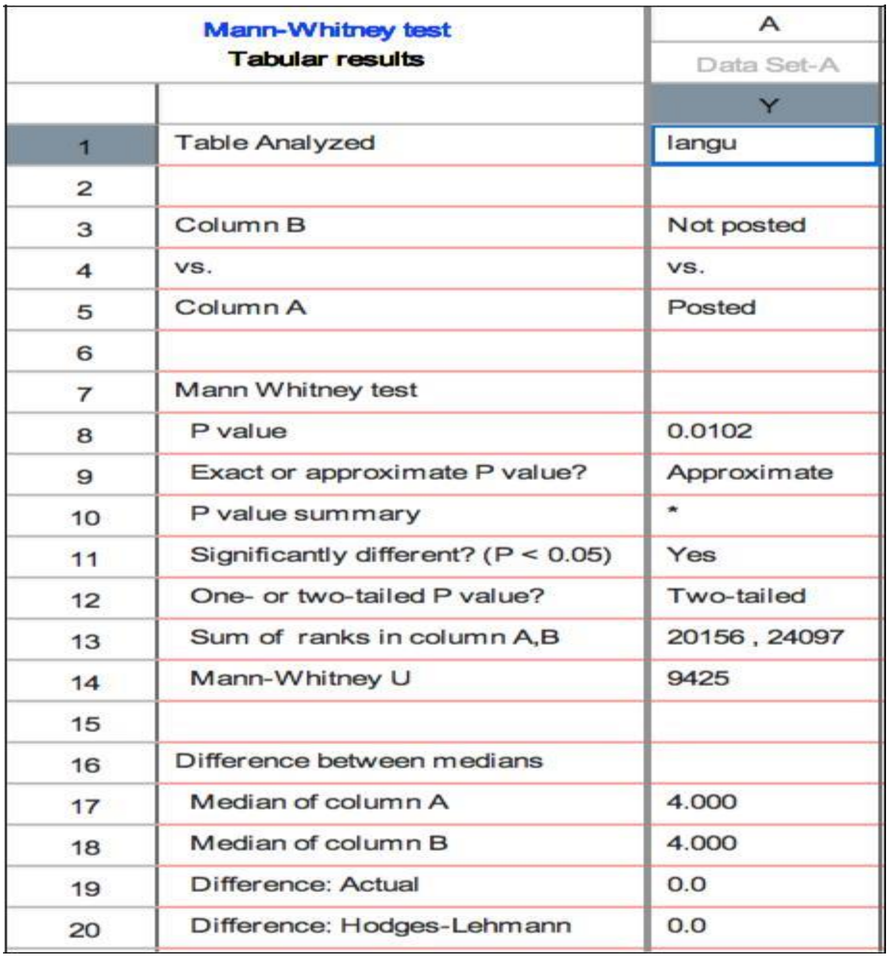


Mann-Whitney test Tabular results		A Data Set-A
		Y
1	Table Analyzed	Punc and Mech
2		
3	Column B	Not posted
4	vs.	vs.
5	Column A	Posted
6		
7	Mann Whitney test	
8	P value	0.0768
9	Exact or approximate P value?	Approximate
10	P value summary	ns
11	Significantly different? (P < 0.05)	No
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	20840 , 24010
14	Mann-Whitney U	9962
15		
16	Difference between medians	
17	Median of column A	4.000
18	Median of column B	4.000
19	Difference: Actual	0.0
20	Difference: Hodges-Lehmann	0.0

3.7 On Language Use

Table 5 shows that there is a significant difference on language use in the compositions produced by the respondents in microblogging and in traditional essay writing. This finding implies that students used more precise words and exerted extra effort when they post their compositions as compared with when they submit their essays directly to their teachers. In an interview with one of the respondents, she stated that she spent longer time writing and revising her work because she is afraid that she would be undermined by her classmates if they would find errors in her essays when they are published online. This finding corroborates with the findings of Muslem et al. (2022) highlighting the positive effect of blogging on ESL students' writing skills, although it is important to note Ozdemir and Aydin's (2015) observation that blogging does not guarantee better writing achievement among EFL writers., and writing instructors should employ effective approaches such as process-based writing to support student blogging activities. In addition, an in-depth analysis of student corpora is needed in order that student vocabulary use is examined, whether they demonstrate what Scarcella (2003) called as major categories of academic vocabulary (i.e., general words used across academic disciplines and in everyday situations; academic words that are common across different disciplines; and technical words found in specific academic fields).

Table 5
Mann-Whitney Tabular Results on Language Use

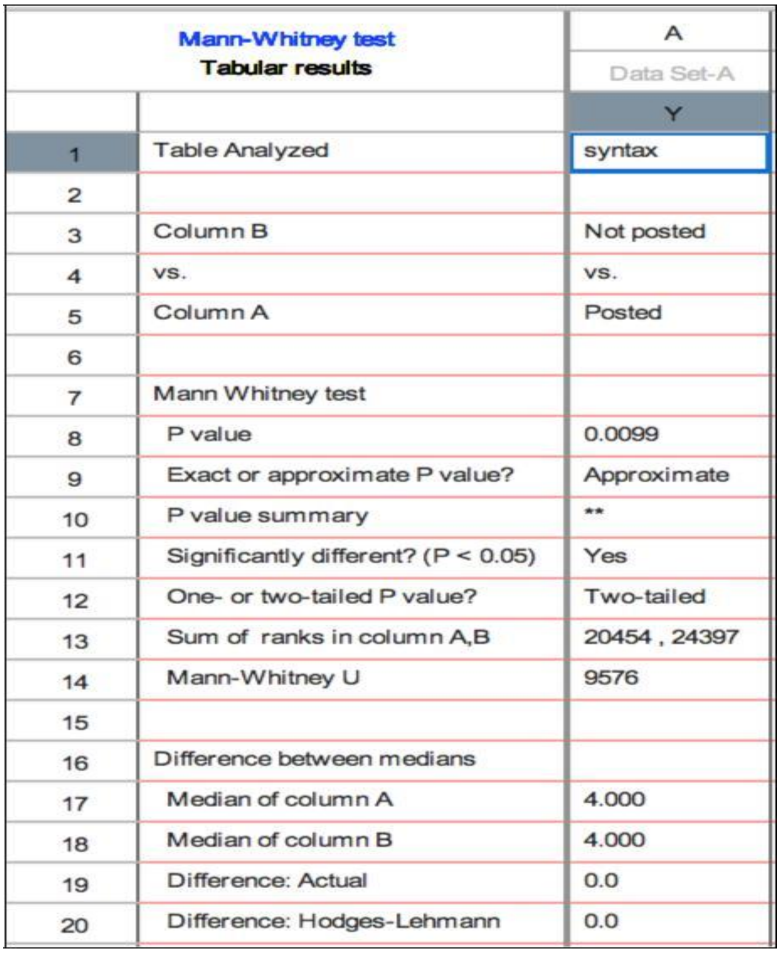


	Mann-Whitney test Tabular results	A Data Set-A
		Y
1	Table Analyzed	langu
2		
3	Column B	Not posted
4	vs.	vs.
5	Column A	Posted
6		
7	Mann Whitney test	
8	P value	0.0102
9	Exact or approximate P value?	Approximate
10	P value summary	*
11	Significantly different? (P < 0.05)	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	20156 , 24097
14	Mann-Whitney U	9425
15		
16	Difference between medians	
17	Median of column A	4.000
18	Median of column B	4.000
19	Difference: Actual	0.0
20	Difference: Hodges-Lehmann	0.0

3.8 On Sentence Structure

Table 6 shows that there is a highly significant difference as regard sentence structure in the compositions produced by students in microblogging than in traditional essay writing. However, while this finding favors blogging over traditional essay writing, there is a need to analyze student corpora in detail, particularly focusing on the syntactic features of student writing, as the grammatical features of academic writing tend to be highly specialized (Biber et al., 2011). In addition, student written outputs in blogging and essay writing require analysis in terms of specific syntactic features which may include embedded phrases (e.g., participial and absolute phrases, and apposition embedded within sentences), complex phrasal structures (e.g., noun and prepositional phrases), and hierarchical structure (e.g., phrase and clause subordination) (Biber et al, 2011).

Table 6
Mann-Whitney Tabular Results on Sentence Structure



Mann-Whitney test Tabular results		A
		Data Set-A
		Y
1	Table Analyzed	syntax
2		
3	Column B	Not posted
4	vs.	vs.
5	Column A	Posted
6		
7	Mann Whitney test	
8	P value	0.0099
9	Exact or approximate P value?	Approximate
10	P value summary	**
11	Significantly different? (P < 0.05)	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	20454 , 24397
14	Mann-Whitney U	9576
15		
16	Difference between medians	
17	Median of column A	4.000
18	Median of column B	4.000
19	Difference: Actual	0.0
20	Difference: Hodges-Lehmann	0.0

Among the six areas that were examined in the data, only the sentence structure is found to have a highly significant difference, while language use is found to be significantly different. Other aspect such as content, organization, and mechanics have shown no significant difference on the two writing strategies employed.

4 Conclusion

The main objective of the study is to examine whether micro-blogging is an effective teaching strategy over the traditional essay writing. Microblogging is when students' essays are posted on the stream section of the Google Classroom where all students in the class could read the written outputs. Traditional essay writing is when students write essays for their teachers to read and mark.

This study suggests that microblogging may be an effective strategy that could be used in ESL writing classes, as students tend to exert more effort to make their compositions better having a wider audience in consideration. This warrants further exploration investigating the possible of impact of audience awareness to the quality of student written outputs, which implicates choice of educators between traditional writing pedagogy or computer-assisted language teaching.

For the study research locale, it is recommended to review the General Education Curriculum and look for possibility where an English course could be offered. In the new curriculum, there is no specific course designed for enhancing the writing or the communication skills of the students; hence, an intensive English course is proposed as an intervention program. It is particularly necessary since English proficiency is seen as a contributing factor in the low performance of university's graduates in various licensure examinations.

This study only delved on microblogging or class blogging but has already shown significant impact on students' writing, albeit on a limited scale. It is therefore recommended that teachers should further explore this strategy as well as the real web blogging as a tool for teaching, where students use online platforms to publish their work. There are several cases of plagiarism in the study. Hence, teachers should be vigilant of this illicit activity as this will defeat the purpose of the teaching and learning process. The University may also review its policy on students' conduct and should not tolerate this kind of practice.

The emergence of online writing in virtual spaces presents opportunities for English language learners to practice their writing skills, as online learning environments afford them to write on-the-go, anytime and anywhere. Such trend opens possibilities for language learning experiences, and our schools should harness this kind of technology and integrate it into our writing classes effectively. As blogging is only a technological tool, teaching EFL/ESL writing should still be informed by time-tested pedagogies and approaches, and executed by teachers with nuanced understanding of the role of technology in the classroom vis-à-vis teaching-learning processes.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Document details – Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? TESOL Quarterly, 45(1), 5-35. https://www.jstor.org/stable/41307614

English Testing Service. (2019). *TOEFL iBT Test Integrated Writing Rubrics*. https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

Gustilo, L. E. (2011). Linguistic features that impact essay scores: A corpus linguistic analysis of ESL writing in three proficiency levels. *3L: The Southeast Asian Journal of English Language Studies*, *17*(1), 55-64.

Gustilo, L. E. (2013). An analysis of writer's performance, resources, and idea generation processes: the case of Filipino engineering students. *Language Testing in Asia*, *3*(1), 1-14.

Mabuan, R. A. (2018). Using blogs in teaching tertiary ESL writing. *English Review: Journal of English Education*, *6*(2), 1. https://doi.org/10.25134/erjee.v6i2.1238

Mohammadi, T., & Mustafa, H. R. (2020). Errors in English writing of ESL/EFL students: A systematic review. *Theory and Practice in Language Studies*, *10*(5), 520-526. http://dx.doi.org/10.17507/tpls.1005.05

Muslem, A., Marhaban, S., Heriansyah, H., & Utama, R. P. (2022). The effects of using blog-assisted language learning (BALL) in improving non-native students' English writing skill in higher education: Does it work? *Journal of Technology and Science Education*, *12*(1), 21-32. https://doi.org/10.3926/jotse.1303

Nordquist, R. (2020, August 27). Understanding Organization in Composition and Speech. https://www.thoughtco.com/organization-composition-and-speech-1691460.

Ozdemir, E., & Aydin, S. (2015). The effects of blogging on EFL writing achievement. *Procedia – Social and Behavioral Sciences*, *199*, 372-380. https://doi.org/10.1016/j.sbspro.2015.07.521

Richards, J. C., & Renandya, W. A. (2002). Methodology in language teaching: An anthology of current practice. Cambridge: Cambridge University Press.

Scarcella, R. (2003). Academic English: A conceptual framework. *The University of California Linguistic Minority Research Institute Technical Report.* https://escholarship.org/uc/item/6pd082d4

Schleicher, A. (2020). *The Impact of COVID-19 on education: Insights from Education at a Glance 2020.* https://www.voced.edu.au/content/ngv%3A87789

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes*. LOT.

Voices in Cyberspace: A Move-Step Analysis of Politics-centered Online News Editorials

Melojie R. Lauron
University of San Carlos
Talamban, Cebu City, Philippines
19104990@usc.edu.ph

Yva Joy B. Dy
University of San Carlos
Talamban, Cebu City, Philippines
19104990@usc.edu.ph

Abstract

This study aimed to analyze the discourse organization of politics-centered news editorials published from 2021 to 2022 by Rappler. Specifically, the paper examined the rhetorical moves and steps that characterize the discourse structures of thirty (30) online news editorials that tackle political issues relevant to the 2022 Philippine elections. Findings revealed that these online news editorials generally followed the two-move pattern in constructing the orientation, exposition, and summation blocks. Regarding the steps employed in the move structures, news editorials demonstrated organizational tendencies reflective of Ho's (2004) Model of General Paper Essay with minor variations in terms of sequencing the steps. Based on these findings, the order of the rhetorical moves and steps varied from one editorial to another as these structural segments were employed. The observation of these rhetorical moves scrutinized how the writers presented their arguments logically to make their claims sound, credible, and agreeable to the target readers.

Keywords: discourse organization, genre analysis, moves and steps, online news editorials

1 Introduction

As an everyday source of information for current events and news, newspaper discourse, specifically editorials, has received particular attention from linguists and any other language study enthusiasts (Tarrayo & Duque, 2011). In light of the 2022 Philippine elections, it is necessary to examine how politics-centered news editorials are structured and organized in orienting and shaping the readers' mindset and stance on this current and significant affair.

According to Anker (2005 as cited in Tarrayo & Duque, 2011), an editorial writer takes position on a controversial issue and gives reasons and supporting evidence to convince readers to accept the position. The role of journalism as a political actor is undeniably evident and the extent to which journalists arrange and organize content to put forward political agenda in news editorials has increasingly piqued the interest of scholars in terms of genre analysis. Recently, the very nature of news consumption has change drastically, especially during the height of pandemic. With the evident shift, the discourse organization of news editorials in the digital space poses an interesting subject for genre-based study.

While there have been existing studies on the analysis of news editorials in various contexts (e.g. Belmonte, 2007; Tarrayo & Duques, 2011; Firmstone, 2019; Syed et al., 2020), to the best of the researchers' knowledge and as evidenced by the relevant literature reviewed, none has been specifically devoted yet in conducting a genre-based analysis on online news editorials that cover political issues in the local context. The present study is intended to fill this research gap. This paper, thus, attempted to make contribution to existing knowledge by analyzing specifically the move and step structures of online news editorials that tackle issues on the 2022 Philippine elections.

The corpus of the present study involves online news editorials published from 2021 to 2022 in Rappler, a Philippine digital news media based in Metro Manila and founded by Filipino journalists. The decision to focus on this online news source is the fact that it is the first all-digital news organization in the country (Ranada, 2018 as cited in Laterno, Magallanes, & Placio, 2019), which has become one of the leading and frequently-read primary sources of people for news consumption. According to Top Websites Ranking for News and Media Publishers in the Philippines updated in 2022, Rappler ranks third in the leading online news brand and first among the all-digital news in the list.

In this regard, the study first attempts to examine the rhetorical move structures that characterize the three blocks of news editorials classified by Ho (2004). Move analysis, as elaborated by Swales (1981, 1990 in Maswana et al., 2015), is a common example in a specific genre-level analysis that examines the communicative purposes of structural segments in a textual discourse. Upton & Cohen (2009) asserts that analysis on moves starts with identifying the sequence of move types and describing the communicative function that each serves in the target genre. With these, the researchers adopt the model of Ho (2004) to describe the overall discourse structure of the political news editorials in relation to the sequence of their move types.

Furthermore, the present study aims to explore the different steps employed in the blocks of online news editorials. Steps for a move refer to the set of rhetorical choices that writers use to realize a specific functional purpose (Yang & Allison, 2003 as cited in Soler-Monreal, 2016). The sub-moves making up the move patterns can be described in terms of interactional roles that correspond to speech acts (Ho, 2004). In relevance, the taxonomy of steps in every move purported by Ho (2004) serves as the framework of this study to determine the sequences of embedded steps in the political news editorials of Rappler.

The findings of this study will contribute to the contemporary understanding on the communicating intentions of online news editorials in tackling political issues, specifically electoral events, as revealed through its structural organization, making it as one of the frequently read primary sources of people for news consumption.

2 Review of Related Literature

2.1 News Editorials on Political Agenda

News editorials are regarded as a complex activity of argumentative writing that takes position on controversial issues and convince readers to accept the position through laying out supporting evidences (Anker, 2005 as cited in Tarrayo & Duque, 2011). Particularly, editorials are written to influence public opinion, encourage critical thinking, and motivate readers to act on an issue. The distinctive role of editorials is that they serve as the collective identity of the editorial board and intend to represent the public voice and opinion of the governing body of a newspaper (Firmstone, 2019).

It is an undisputed fact that media, specifically in electronic forms, have provided people a forum for political expression. El Baff et al. (2020) states that news editorials “argue about political issues in order to challenge or reinforce the stance of readers with different ideologies” (p. 3154). At present, demand for more transparency in politics is increasingly growing. The active engagement of citizens to political topics shows how they form their opinions about relevant issues through paying opinion to the information supplied by the media.

2.2 Moves and Steps in News Editorials

Previous studies have begun to examine how information presented in news editorials are structured and organized by analyzing the moves and steps. Tongsibsong (2012) conducted a genre analysis of English editorials regarding hard news in broadsheet and tabloid newspapers focusing on move analysis using the frameworks of move analysis introduced by Swales (1990) and Bhatia (1993). Findings revealed that there is one common move structure in the broadsheet newspaper’s editorials and two common move structures in tabloid newspaper editorials. Six similar features of the editorials in broadsheet and tabloid newspapers were also found. With the same intent, the present study intended to analyze the moves and steps employed in the discourse organization of online news editorials.

Another study was done by Orosa, Garcia, and Santorum (2013) who conducted an in- depth content analysis on the features and degrees of adaptation of the editorials of five newspapers from different European countries to the communicative online environment. The traditional and online editorials were compared and contrasted according to roles, main objectives, structures, text types, use of hypertext and multimedia, reader's participation, and updating. The findings revealed that the online editorials maintain the traditional content and format features of printed editorials, and shows little adaptation to the digital platform. Unlike Orosa et al.'s (2013) study, this paper focused only on examining online news editorials taken from Rappler, which mainly publishes digital news in the Philippines.

Tarrayo and Duque (2011) described the discourse structure and textual metadiscourse in Philippine newspaper editorials. The research utilized Ho's (2004) model of a possible discourse structure of a General Paper essay in analyzing a total of 24 editorials published in 2010 by two leading Philippine broadsheets. Results of the study indicated that the articles followed the two-move pattern in each block. Like any other argumentative essays, the editorials are structured in a way that the orientation block introduces the writer's claim on an issue; the exposition block expands the discussion of this claim; and the summation block provides a summary of the writer's primary point and a recommendation or solution.

In another study conducted by Gapas (2016), Ho's (2004) model of General Paper (GP) essay was also adopted as a framework to formally investigate the discourse organization of opinion articles found in university-level newspapers. Findings of the study revealed that the articles observed the two-move pattern in the orientation, exposition, and summation blocks and employed two frequently used specific steps in each block.

Similarly, the present study also adopts Ho's (2004) framework to investigate the discourse organization of political news editorials from online news sources. Specifically, the model is utilized to examine the specific moves and steps that were employed in presenting political issues through the editorials. Although the present study used Ho's (2004) model, this paper mainly focused on thirty (30) news editorials published in Rappler.

By and large, these empirical studies serve their purpose as substantial references to the present study. As observed, most of the reviewed studies were only focused on the discourse structure of printed news editorials and there was no study that analyzed online editorial articles in the local context. Hence, there is a need to fill the research gap through analyzing the discourse organization of online news editorials that tackle political agenda in light of the 2022 Philippine elections.

3 Theoretical Background

Ho's (2004) Model of a Possible Discourse Structure for a General Paper (GP) Essay serves as the framework for the current study.

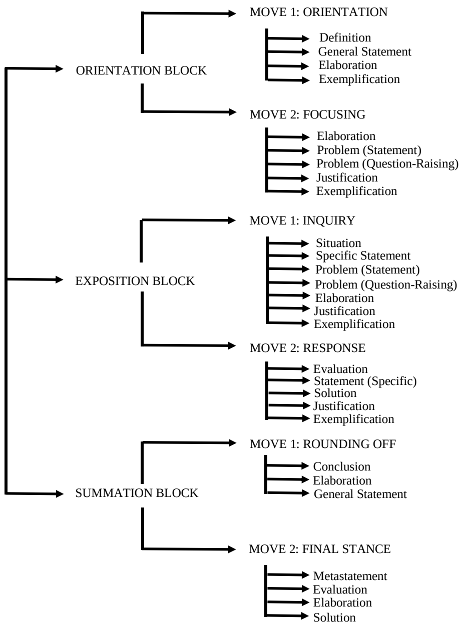


Figure 1. The model of a general paper essay by Ho (2004)

The characterization of the online editorial texts as a discourse unit is done by adopting Ho's (2004) Model of a General Paper Essay. Ho (2004) described the possible text structure and organization patterns that characterize the General Paper (GP) essay, in which she proposed three blocks in this model, namely the Orientation, Exposition, and Summation block. Each block is comprised of a two-move pattern that consists accordingly of specific steps.

Based on the literature reviewed, local printed opinion articles and newspaper editorials have been examined through the framework of Ho (2004), which marks the model's observance to the journalistic principles of editorial writing. Examining the corpus of the present study using this model reveals how the general components of the editorial's macrostructure – orientation, exposition, and summation blocks – are structured and organized in tackling politics-centered issues.

4 Research Objective and Questions

This study aimed to conduct a genre-based analysis on the discourse organization of politics-centered news editorials published from 2021 to 2022 by Rappler, a Philippine-based digital news media, through examining the rhetorical moves and steps employed in online editorials that tackle political issues in light of the 2022 Philippine elections.

Specifically, it sought to answer the following questions:

1. What rhetorical moves as proposed by Ho (2004) characterize the discourse structures of Rappler's news editorials?; and
2. What specific steps are employed in the move structures of the orientation, exposition, and summation blocks of the online news editorials?

5 Methodology

This study aimed to conduct a genre-based analysis on the discourse organization of politics-centered news editorials published from 2021 to 2022 by Rappler, a Philippine-based digital news media, through examining the rhetorical moves and steps employed in online editorials that tackle political issues in light of the 2022 Philippine elections. Structural interpretation of the rhetorical moves and steps from printed and online news sources are based on the model of Ho (2004).

The research study employed a qualitative-quantitative design in analyzing thirty (30) editorial news articles published from 2021 to 2022 by Rappler. The quantitative design of the study utilized frequency counts of the moves and steps identified by Ho (2004) that were evident in the news editorials. As for the qualitative design, the research employed rhetorical analysis to describe the discourse organization of news editorials published in online news sites based on the move structures and steps employed.

A total of thirty (30) genre samples of news editorials were carefully analyzed to examine the discourse organization of the editorials. Selection of the genre samples for analysis was based on the relevance of the issues to the 2022 Philippine elections. After the initial steps, the news editorials collected were encoded and reformatted using word-processing software such as Microsoft Word. Each article was first analyzed, divided, and classified according to the three blocks presented in the model of Ho (2004). Afterwards, macro data analyses were conducted, where the rhetorical moves and the specifics steps used in the political news editorials were examined based on the General Paper Essay Model of Ho (2004).

After the discourse blocks, moves, and steps were classified based on Ho's (2004) model, the occurrences of the rhetorical moves and steps in every block were counted, tallied, and tabulated on a frequency table. Whereas the organizational moves were counted and tallied based on whether they occurred in each block for every online news editorial, the steps were counted based on their occurrence in every rhetorical move.

To answer the first subproblem, the researchers counted the rhetorical moves that characterize the discourse structure of online news editorials. Results were tallied manually and were tabulated for the frequencies and percentage of the moves. A rhetorical analysis was utilized for further genre-based analysis.

As for the second subproblem, the news editorials were examined by determining the steps that were utilized in every move as identified by Ho (2004). The researcher tallied and tabulated the results to show the frequency and percentage of the steps evident in the editorial articles.

After the presentation of each move and step, the researchers extracted specific samples from these online news editorials to illustrate the most prevalent move and step structures in the discourse organization of news editorials that tackle issues regarding the country's upcoming national elections.

6 Results and Discussion

This section presents the key results of the study with reference to its aims. The sequence of the discussion is as follows: (1) the rhetorical moves that characterize the discourse structure of online news editorials; and (2) specific steps employed in the move structures of each block of online news editorials.

6.1 Moves Employed in Rappler's Online News Editorials

The data in the following table present the frequency and percentage distributions of move patterns employed in the political news editorials from Rappler news.

Move	F	%
Two-move pattern (i.e., Orientation block: Orientation/Focusing; Exposition block: Inquiry/Response; Summation block: Rounding off/Final stance	27	90
Non-two-move pattern	3	10
Total	30	100

Table 1: Move patterns utilized in online news editorials

The findings in Table 1 suggest that the organization of online news editorials that tackle political issues in the country generally observes the two-move pattern in the orientation, exposition, and summation blocks classified in the study of Ho (2004). This indicates that the presentation of political issues in online news editorials may be structured in a manner similar to typical argumentative essays. Specifically, the news editorials were organized as follows: the orientation block provides an overview or a background of the topic, including the news media's position on the political issue; the exposition block elaborates the general claim and gives an in-depth discussion of the issue; and the summation block synthesizes the ideas presented and discussed.

Steps	F	%
Orientation		
Orientation	30	17.44
Focusing	27	15.70
Exposition		
Inquiry	30	17.44
Response	28	16.27
Summation		
Rounding off	30	17.44
Final Stance	27	15.70
Total	172	100

Table 2: Move structures utilized in online news editorials

The findings in Table 1 suggest that the organization of online news editorials that tackle political issues in the country generally Table 2 above shows the rhetorical moves utilized in the news editorials from Rappler. Based on these findings, it can be inferred that the online articles frequently employ the first move in every block. Examining each, orientation, inquiry, and rounding off moves are consistently utilized in the orientation, exposition, summation blocks of all thirty (30) online news editorials, respectively.

First, in the orientation block, the use of the "Orientation" move indicates that in presenting the political issues in online news editorials, writers commonly begin by providing the background and context of the topic. This is followed by the "Focusing" move where the writer focuses on the discussion of a specific issue. Second, in the exposition block, the consistent use of "Inquiry" move suggests that online editorials tackle political issues by directly providing specific details that support the main argument. This is, then, followed by the "Response" move that allows the writer to develop a more in-depth discussion of the main ideas in the editorials, including their input on the issue. Lastly, "Rounding Off" move in the summation block is where the writer restates the thesis and provides conclusion based on the main ideas pointed out in the political issues covered in the articles. As for the "Final stance" move, this is used as a means of evaluating and further elaborating the ideas stated in the orientation block.

6.2 Steps Employed in Rappler's Online News Editorials

The data in the following table present the frequency and percentage distributions of the steps utilized in the orientation, exposition, and summation blocks of political news editorials taken from Rappler news.

Orientation Block

Steps	F	%
General statement	30	37.97
Elaboration	23	29.11
Problem (Statement)	8	10.13
Problem (Question-raising)	7	8.86
Justification	5	6.33
Exemplification	4	5.06
Definition	2	2.53
Total	79	100

Table 3: Steps employed in the orientation block of online news editorials

Findings in Table 3 reveal that online news editorials use the general statement and elaboration steps more frequently than the other steps in presenting political issues. The above results agree with the findings of Gapas (2016) in his study regarding the discourse organization of Philippine university newspaper opinion columns. Introductions in the editorial articles must provide a general background of the presented argument to sustain the readers' attention and interest (Tan, 2003). Moreover, as emphasized by Brizee & Tardiff (2014 as cited in Gapas 2016), general statements consist of thesis statements that may contain the claims, points of views, interpretations, and cause-and-effect statements that the writers aim to convey. The use of the elaboration step is also evident in the orientation block of the news editorials, in which this step provides details, particulars, and any other elaborations of the preceding statement (Ho,2004).

Aside from the aforementioned steps, the table also reveals the use of problem- statements and problem-questioning steps in online news editorials. Ho (2004) posits that these steps present aspects of a situation that require a response from the readers. It is through the use of either statement or explicit question form that the steps can be actualized. As writers provide general statements, they resort to the use of justification step to present reasons in support of the stated idea, and exemplification step to provide specific examples and concrete data and statistics. In addition, the definition step, which garnered the least frequency counts, is used in online news editorials to give context to an issue by explaining the meaning of a concept, particularly the unfamiliar terms.

Exposition Block

Steps	F	%
Elaboration	85	20.83
Specific Statement	81	19.85
Problem (Question-raising)	70	17.16
Justification	58	14.22
Situation	31	7.60
Exemplification	26	6.37
Problem (Statement)	22	5.39
Evaluation	21	5.15
Solution	14	3.43
Total	408	100

Table 4: Steps employed in the exposition block of online news editorials

Upon examining the data presented, it can be inferred that the total frequency of steps in the exposition block of online news editorials has the highest number of occurrences compared to the other blocks. This can be attributed to the fact that sentence lengths in the exposition block vary, which involve the multiple uses of moves and steps to establish the particular details of the political issues and arguments tackled. This observation gives a nod to the assertion of Ho (2004) that among all the three blocks, exposition is the least predictable in terms of its move-step structure.

From these findings, it can be observed that elaboration and specific statement are two of the most frequently used steps in the online news editorials. In tackling political issues, these news editorials present specific statements to state a particular claim or proposition (Ho, 2004) and construct specific arguments in support of the main argument (Tan, 2003). These arguments, then, are developed and strengthened through the use of the elaboration step to provide detailed explanations of the preceding statements.

Moreover, other frequent steps that occurred in this block are problem (question-raising) and justification. After presenting a particular argument, the writers pose an explicit question to raise a problem and acquire a response from the readers. Given that the "response" move in this block primarily contains the input of the writers on the topic, writers justify the arguments presented through providing reasons in support of the given statement. They also employ the situation step to provide background information by presenting facts and circumstances, particularly on the discussion of political issues. With the use of exemplification step, editorial writers are able to illustrate aspects of these propositions and assertions through giving concrete data and specific examples to compel readers to accept the claims.

Summation Block

Steps	F	%
Conclusion	24	25
Elaboration	24	25
General statement	19	19.79
Evaluation	15	15.63
Solution	14	14.58
Metastatement	0	0
Total	96	100

Table 5: Steps employed in the summation block of online news editorials

Table 5 presents frequency and percentage distributions of the steps used in the summation block. The summation block utilizes the "rounding off" and "final stance" moves. Data show that among the summation steps, the most frequently used are elaboration and conclusion. Contrary to the findings of Gapas (2016) where not many writers opted to provide a direct conclusion, the genre samples of editorials in the present study frequently use conclusion to reiterate the ideas in the preceding statements, and summarize the main points of the issue.

Moreover, the general statement in online news editorials is utilized to state in broad terms the proposition presented in the summation block. The use of evaluation and solution steps is also evident, where news editorials evaluate the circumstances and put forward recommendations and proposals as to how the problem could be solved. As stated by Tan (2003), closing paragraphs must leave a strong impression to the readers. This can be achieved by integrating evaluations, possible solutions, and judgements. With reference to the present findings, this assertion is observed in the online news editorials that tackle politics-related issues.

Based on the sequences of steps in the online new editorials that tackle political issues in the country, it can be noted that some editorial articles do not exactly follow the structural order of Ho's model as these steps are employed depending on the purposes and intentions that the writers hope to convey in the editorials. In a similar sense, the order of the rhetorical moves also varies from one editorial article to another, in which some online news editorials are characterized with two-move patterns, while others follow a non-two-move pattern.

By and large, observation of the moves and steps give readers a logical framework of the editorial writers' ideas that will allow them to weigh carefully the information received. This allows news consumers to carefully examine how the writers present concrete pieces of evidence logically to prove their claim and make their arguments sound, credible, and agreeable to the target readers. All of these can potentially influence people's future decisions and actions, especially that the 2022 election is approaching and the readers, who are citizens of this country, must be wise in choosing the deserving leaders.

Conclusion

Based on the findings, these are the conclusions of the study:

1. The discourse structures of politics-centered news editorials from the online news site, Rappler¸ generally follow the two-move patterns in constructing the orientation, exposition, and summation blocks. These patterns are only apparent to some extent as there are few news editorials that do not observe the two-move pattern. Regardless, all news editorials consistently employ the first moves, namely the orientation, inquiry, and rounding off, in every block. These suggest that online news editorials tackle political issues in a manner that begins with establishing the background and context of the topic, presenting, then, specific details to support the main arguments, and synthesizing the ideas and main points raised to ensure effective communication of the topics and stances to the readers.
2. With regard to the steps employed in the move structures of Rappler's news editorials, the orientation block frequently used the general statement and elaboration steps; the exposition block revealed the frequent occurrences of elaboration and specific statement steps, and the summation block

frequently utilized the conclusion and elaboration steps in synthesizing the main ideas presented. These steps demonstrated organizational tendencies reflective of Ho’s (2004) model with minor variations in terms of the sequencing of steps. In other words, it can be noted that not all online news editorials adhere exactly to the structural order of Ho’s model as evident in the varied sequences of the steps employed. This implies that the organization of rhetorical steps in online news editorials depends on the objectives and purposes of the writers.

Recommendations

Based on the findings and conclusions, the following are recommended:

1. Given the limited number of samples on this genre-based analysis, future studies with larger samples or corpus of online news editorials can be undertaken to arrive at more conclusive results.
2. Online news editorials can serve as authentic examples of argumentative texts that can be used in pedagogy to aid learners in analyzing discourse structures, in evaluating facts and viewpoints, and in presenting their own arguments concerning current news and affairs.
3. Ho’s (2004) Model of General Paper (GP) Essay should be introduced not only to students, but also to writers in general. The moves and steps in the model can be a helpful guide to them in constructing and organizing written discourses, and maintaining the coherence of the ideas presented.
4. The prevalent rhetorical move and step patterns in online news editorials can be employed or enhanced by writers in the process of writing essays to provide sound claims and logical arguments to the target readers.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Belmonte, I.A. (2007). Newspaper editorials and comment articles: A “cinderella” genre. Revista Electronica de Linguistica Aplicada, 1: 1-9. Retrieved from https://www.researchgate.net/publication/28205752_Newspaper_editorials_and_comment_articles_a_cinderella_genre.

El Baff, R., Wachsmuth, H., Al Khatib, K., Stein, B. (2020). Analyzing the persuasive effect of style in news editorial argumentation. Association for computational linguistics, 1: 3154- 3160. doi: http://dx.doi.org/10.18653/v1/2020.acl-main.287

Firmstone, J. (2019). Editorial journalism and newspapers’ editorial opinions. Oxford Research Encyclopedia of Communication, 1: 1-22. doi: 10.1093/acrefore/9780190228613.013.803.

Gapas, W. G. (2016). The discourse organization of Philippine newspaper opinion columns. Asian Journal of English Language Studies, 4: 34-53. Retrieved from https://ajels.ust.edu.ph/wpcontent/uploads/2018/09/3-The-discourse-organization-of- Philippine-university-newspaper-opinion-columns.pdf.

Ho, C. (2004). Discourse features and strategies in students’ argumentative writing at pre- university level. ACELT journal, 8: 3-10.

Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. Ampersand, 2: 1-11. doi: 10.1016/j.amper.2014.12.002.

Lanterno, M., Magallanes, M., & Placio, M. (2019). The Credibility of Rappler as Perceived by Communication Students from Selected Universities in Metro Manila: Academic Year 2018-2019. (Unpublished thesis). Polytechnic University of the Philippines, Sta. Mesa, Manila.

Newman, N. D. Levy, F. & Nielsen, R.K. (2019). Reuters Institute Digital News report 2019 (Report). Oxford: Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019- 06/DNR_2019_FINAL_0.pdf.

Orosa, B.G., Garcia, X.L., Santorum, S.G. (2013). Analysis of the adaption of the editorials of five newspapers from different European countries to the online environment. Revista Latina de Communicacion Social, 68 : 485-501. doi: 10.4185/RLCS-2013-986en

Soler-monreal, C. (2015). A move-step analysis of the concluding chapters in Computer science Phd theses. Revista de la Asociacion Europea de Lenguas para Fines Especificos, 32 : 105-132. Retrieved from https://www.redalyc.org/journal/2870/287048507006/html/.

Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

Syed, S. El Baff, R., Kiesel, J. Khatib, K. A., Stein, B., Potthast, M. (2020). News editorials: Towards summarizing long argumentative texts. Proceedings of the 28th International Conference on Computational Linguistics, 5384-5396. doi: 10.18653/v1/2020.coling- main.470.

Tan, K. (2003). Constructing the argumentative essay. Successful Learning. Retrieved From http://www.cdtl.nus.edu.sg/success/sl29.htm.

Tarrayo, V. & Duque, M.C. (2011). Arguing in l2: discourse structure and textual metadiscourse in Philippine newspaper editorials. Journal on English language Teaching, 1(4): 11-24. Retrieved from https://eric.ed.gov/?id=EJ1071061.

Tongsibsong, J. (2012). A genre analysis in English editorials regarding hard news in broadsheet and tabloid newspapers. National institute of development Administration: School of language and communication. Retrieved from http://libdcms.nida.ac.th/thesis6/2012/b180478.pdf .

Top Websites Ranking (2022). Top News and Media Publishers in Philippines Ranking Analysis for August 2022. Retrieved from https://www.similarweb.com/top-websites/philippines/category/news-and-media/

Upton, T. & Cohen, M. (2009). An approach to corpus-based discourse analysis: The move analysis as example. Sage journals, 11(5): 585-605. 10.1177%2F1461445609341006.

Appendices

Appendix A. News Editorial Articles

Rappler News Editorials from https://www.rappler.com/

"2021 is when we act and build, or bust" (January 4, 2021)
"Why is the government afraid of community pantry" (April 6, 2021)
"China and Duterte's roaring defense" (May 10, 2021)
"The boxing in the Senate and the circus 2022" (May 31, 2021)
"Politics first before brownout, before town" (June 7, 2021)
"Duque, Lao, and Duterte's crackdown" (August 23, 2021)
"Get rid of false heroes" (August 30, 2021)
"The promoters of money heist in the pandemic" (September 13, 2021)
"Pharmally hits the pavement" (September 27, 2021)
"Justice is coming" (September 20, 2021) "Mr. Duterte, that's fine!" (October 4, 2021)
"When the Ombudsman weaponizes SALNs against the public" (October 25, 2021)
"The deluge of lies" (November 1, 2021)
"Criminal or immoral: Accepting money during election period" (November 3, 2021)
"Why the 2022 presidential candidates must have a global outlook" (November 4, 2021)
"Substitution in aid of gangster politics" (November 15, 2021)
"This is the era of fake hero" (November 29, 2021)
"We have one, big democracy project. What to do?" (December 13, 2021)
"Outshining the tyrant" (December 27, 2021)
"2022, the year to be political" (December 28, 2021)
"P203-billion debt of the Marcos family, charge!" (March 25, 2022)
"New Year call to Comelec: Follow in the footsteps of 35 tabulators" (January 3, 2022)
"Do not be infected with Marcos Jr.'s Zoomicron" (January 19, 2022)
"Evicting justice" (January 24, 2022)
"Where are you headed?" (February 7, 2022)
"Is this the last EDSA?" (February 21, 2022)
"Be smart in voting, oil fuel crisis is here" (March 14, 2022)
"Our democracy is not a family business" (March 28, 2022)
"Lorraine Badoy, desecrated truth" (April 4, 2022)
"Let us reclaim public space" (May 7, 2022)

Probable Human Knowledge Elicitation for Efficient Human-in-the-Loop Text Classification without Any Label-Related Data

Osamu Saisho
NTT / Tokyo, Japan
osamu.saisho.vm@hco.ntt.co.jp

Abstract

This paper presents a new human-in-the-loop text classification framework with no labeled data incorporating weak supervision and Bayesian active learning with a semantic clustering constraint. Labeling function (LF)-based weak supervision is a promising method to alleviate the huge labeling cost in supervised learning. However, most previous studies relied on the impractical assumption of sufficient and well-designed pre-built LFs, even though they also require a heavy workload to implement. The proposed framework is the first intended to substantially reduce this workload by a human-in-the-loop approach from scratch. Specifically, it interactively suggests data samples to help humans implement an LF considering the actual human workload. Its superiority is achieved by Bayesian active learning with a semantic clustering constraint for not only enhancing labeling utility but also reducing human workload. The results of the user studies with 16 volunteers demonstrate that the proposed method can achieve higher performance with fewer LFs and help humans implement more accurate LFs in a shorter work time than conventional methods.

1 Introduction

Supervised learning assuming the availability of massive labeled data has been put into practical use, but this assumption has also been the main obstacle to expanding its use. To overcome this obstacle, weak supervision (WS) has gained much attention as a promising method. It is incomplete, inexact, or inaccurate supervision as an alternative to strong supervision such as a fully labeled dataset (Zhou, 2017). Specifically, labeling function (LF) is a representative approach used in both academia and industry (Bach et al., 2019; Fries et al., 2019). LF is a human-defined function based on heuristics, collective knowledge, or implicit rules for indirect labeling instead of manual direct labeling. Its main strengths are its high performance and robustness to task changes (Ratner et al., 2016; Bach et al., 2017).

Even though LF-based WS alleviates the huge and redundant labeling cost, it usually depends on sufficient and well-designed pre-built rules and meta-rules for LFs. When implementing LFs in the wild from scratch, the difference in the required workload and the potential benefits on each LF must be considered. Thus, there are two issues left to be solved simultaneously for the practical LF-based WS framework. First, only high-performance LFs should be implemented. Second, humans should be able to implement more accurate LFs with a lighter workload, that is, in a shorter work time.

The most appropriate and pragmatic approach to solve both issues is a human-in-the-loop approach. Especially, active learning (AL) (Settles, 2009) is a key method to solve the first issue. A few recent studies have attempted to develop an AL-like framework for WS (Wang et al., 2019; Saisho et al., 2021). These methods try implementing effective LFs interactively, but they disregard the workload for each LF. In addition, their effectiveness has been verified only in simulation settings under unrealistic assumptions such as “selecting” the best one among the “pre-defined” candidates for each loop. From the viewpoint of practical WS in the wild from scratch, the superiority in LF should be determined on the basis of the work time to design it, the voting coverage of the applicable data samples, and the accuracy of the voting, other than the labeling utility in AL.

The research objective is to reduce the substantial workload of humans while achieving high performance in the practical WS framework in the wild from scratch. The main contribution of this paper is the proposal and verification of a practical human-in-the-loop WS framework by Bayesian AL with a semantic clustering constraint. In each loop, this framework suggests data samples that not only have high utility for labeling but also elicit probable human knowledge to help humans evoke common labeling rules for a new LF with a lighter workload. Especially, the data samples are suggested on the basis of utility scores calculated by batchBALD acquisition function under the constraints of the batch being formed by data samples belonging to an identical semantic cluster. This constraint is the key to filling the gap in the practice of a human-in-the-loop WS. The results of the user studies with 16 volunteers verified the high performance of the classifier trained by using a small number of LFs, the short work time required to implement LFs, and the high accuracy of the LFs' votings. As an adjunct to the above contribution, this paper also shows an empirical validation method in human-in-the-loop WS that uses LFs as a substitute for ground truth, without any additional workload or implicit label leakage.

2 Related work

Towards the practical LF-based WS, the reduction of LF implementation cost has gained considerable attention in the past few years. Some researchers have proposed fully automated methods by transfer learning and few-shot learning approaches (Varma and Ré, 2018; Das et al., 2020). Their insights are significant, but they have also eliminated positive aspects of WS in practical use. They incur the uninterpretability because of the dependency on the implicit domain similarity and the distribution of the few-shot data (Raghu et al., 2019; Chen et al., 2019). On the other hand, a human-in-the-loop approach will work well with LF-based WS because of their analogous advantages if humans can implement LFs from their knowledge with light workload. Transparency and fairness are the primary reasons that human-in-the-loop is attracting attention (Holzinger, 2016; Li, 2017; Lertvittayakumjorn et al., 2020). Especially, AL is the representative human-in-the-loop approach toward reducing labeling costs.

Existing methods to combine WS and AL can be categorized into four types. The first type is "AL with WS" (Nashaat et al., 2018; Nashaat et al., 2020; Biegel et al., 2021). This type trains a classifier using "pre-built" WS and then improves its classification performance by direct labeling in AL. The second type is "WS with AL" (Qian et al., 2020; Gonsior et al., 2020). This type attaches labels directly by AL and augments the labeled dataset by "pre-built" WS. These two types can lighten somewhat humans' tasks in the conventional AL, but they cannot change the redundancy of the tasks. The third type is "AL for automated WS" (Kartchner et al., 2020; Boecking et al., 2021). This type is an AL framework that automatically generates WS on the basis of "pre-built" meta patterns and asks humans to judge whether to accept them or not. It can eliminate the redundant labeling cost, but their performance depends on the pre-built resources such as seed rules and structure information as the automated WS above. The most significant problem common to these three types in practice is that they all still depend on the existence of unrealistic pre-built resources. Even though they are human-in-the-loop approaches, preparing such well-designed meta resources without any support before the loops is difficult and unrealistic in the wild.

The last type is "AL for WS from scratch" (Wang et al., 2019; Saisho et al., 2021). This type is an AL framework that helps humans implement WS by suggesting a few data samples that have the highest labeling utility. Only this type is pragmatic in the wild because it does not require any presupposed resources and humans can implement WS in any format. It can reduce the number of LFs to be manually implemented, but this is still not enough for practical use in the wild. It is not human-friendly because it does not consider any characteristic of human-defined LFs. First, each LF requires a very different amount of workload to implement. Furthermore, both the coverage and the accuracy of LFs need to be enhanced in the case of WS, where there is no oracle

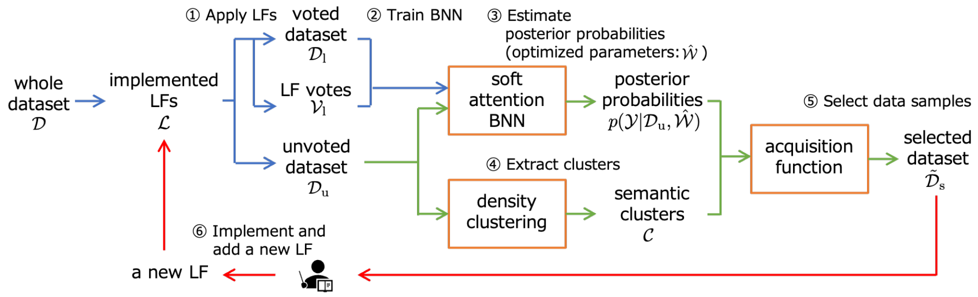

① Apply LFs
voted dataset $\mathcal{D}_\mathrm{l}$
② Train BNN
③ Estimate posterior probabilities (optimized parameters: $\hat{\mathcal{W}}$)
whole dataset $\mathcal{D}$
implemented LFs $\mathcal{L}$
LF votes $\mathcal{V}_\mathrm{l}$
soft attention BNN
posterior probabilities $p(\mathcal{Y}|\mathcal{D}_\mathrm{u}, \hat{\mathcal{W}})$
⑤ Select data samples
unvoted dataset $\mathcal{D}_\mathrm{u}$
acquisition function
selected dataset $\tilde{\mathcal{D}}_\mathrm{s}$
④ Extract clusters
density clustering
semantic clusters $\mathcal{C}$
⑥ Implement and add a new LF
a new LF

Figure 1: Overall framework: a few data samples are selected by batchBALD under the clustering constraint and then a human implements a new LF referring to them in each loop (performed in the order of blue, green and red).

that always attaches correct labels (Dasgupta, 2005; Ipeirotis et al., 2014). As a result, the superiority of the existing methods is verified only in simulation settings, i.e. human-in-the-loop without human settings. Therefore, the proposed framework is the first practical method of AL for WS from scratch intended to substantially reduce human workload.

In addition, AL itself has faced criticisms for practical use (Siddhant and Lipton, 2018; Atighehchian et al., 2020). The main concerns are inaccuracy of estimated uncertainty and implicit label leakage. To overcome the first concern, Bayesian AL such as Bayesian AL by Disagreement (BALD) (Houlsby et al., 2011) has often been used to estimate the uncertainty more accurately than the traditional uncertainty-based methods (Yuan et al., 2019). The proposed framework utilizes batchBALD with consistent Monte Carlo dropout (cMCdropout) (Kirsch et al., 2019) as an acquisition function in AL, which is an extension of BALD to select multiple data samples in a single loop. The model uncertainty is calculated from the variance in the output of multiple runs with different dropout masks (Gal and Ghahramani, 2016) but fixed during each epoch in cMCdropout. The maximum batchBALD score can be calculated by greedy approximation thanks to the submodularity of the mutual information. Label leakage, the second concern, means that fully labeled validation datasets are implicitly used for hyperparameter optimization or validation during training, even though they are unavailable in a practical setting. This concern can also be found in some WS studies. The proposed framework entirely avoids making any implicit label leakage by using LFs as a substitute for ground truth in the validation step.

3 Framework

The proposed framework is a new method of AL for WS from scratch, so the computer suggests a few prioritized data samples, and humans implement an LF in each loop. What is important is how to select the data samples to help humans design a more effective and probable LF with a lighter workload. To ensure the effectiveness, the proposed framework has a soft attention Bayesian neural network (BNN) derived from Ren et al. (2020) that both estimates Bayesian uncertainty via cMCdropout and reduces label noise of WS by attention mechanisms. In addition, to reduce the human workload, the proposed framework has density clustering to aggregate data samples having common semantic features and adds a clustering constraint to the acquisition function for considering the aggregation.

This section describes the proposed framework by dividing it into six steps in each loop as shown in Figure 1 ① - ⑥ for easy understanding. These six steps are executed in order in each loop. Note that the second and third step set and the fourth step can be executed in parallel because there is no dependency on their input data. The proposed framework is designed for cooperation between humans and computers, so each data sample $\boldsymbol{d}$ in the whole unlabeled dataset $\mathcal{D}$ is assumed to have two representation styles: (1) a human-recognizable representation such as a raw text; and (2) a feature vector representation by embedding or feature extraction. Also, let $\mathcal{Y}$ and $\mathrm{n}(\cdot)$ denote the class label set and the number of elements in a set, respectively.

The first step is applying LFs. Each LF in the set of implemented LFs $\mathcal{L}$ receives $\boldsymbol{d}$. The votes for $\boldsymbol{d}$ by all LFs in $\mathcal{L}$ are combined into a voting result vector $\boldsymbol{v} \in \{-1, 1, 2, \ldots, \mathrm{n}(\mathcal{Y})\}^{\mathrm{n}(\mathcal{L})}$. If the l-th compo-

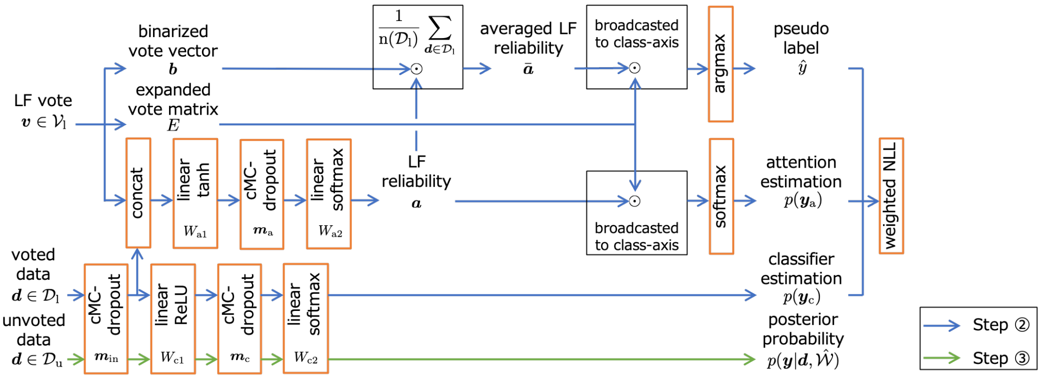

binarized vote vector $\boldsymbol{b}$
expanded vote matrix E
LF vote $\boldsymbol{v} \in \mathcal{V}_\mathrm{l}$
$\frac{1}{\mathrm{n}(\mathcal{D}_\mathrm{l})} \sum_{\boldsymbol{d} \in \mathcal{D}_\mathrm{l}}$
averaged LF reliability $\bar{\boldsymbol{a}}$
broadcasted to class-axis
argmax
pseudo label $\hat{y}$
concat
linear tanh W_a1
cMC-dropout $\boldsymbol{m}_\mathrm{a}$
linear softmax W_a2
LF reliability $\boldsymbol{a}$
broadcasted to class-axis
softmax
attention estimation $p(\boldsymbol{y}_\mathrm{a})$
weighted NLL
voted data $\boldsymbol{d} \in \mathcal{D}_\mathrm{l}$
unvoted data $\boldsymbol{d} \in \mathcal{D}_\mathrm{u}$
cMC-dropout $\boldsymbol{m}_\mathrm{in}$
linear ReLU W_c1
cMC-dropout $\boldsymbol{m}_\mathrm{c}$
linear softmax W_c2
classifier estimation $p(\boldsymbol{y}_\mathrm{c})$
posterior probability $p(\boldsymbol{y}|\boldsymbol{d}, \hat{\mathcal{W}})$
Step ②
Step ③

Figure 2: Structure of the soft attention BNN: the posterior probability is calculated with Bayesian uncertainty thanks to cMCdropout. To denoise the WS, the LF reliability is calculated for each data sample concatenated with the LF votes and then weighted majority voting is performed using the LF reliability.

nent of $\boldsymbol{v}$ is -1, it means that the l-th LF has abstained from voting. After the voting, $\mathcal{D}$ is divided into a voted dataset $\mathcal{D}_\mathrm{l}$ and an unvoted dataset $\mathcal{D}_\mathrm{u}$ depending on the voting results. If $\boldsymbol{d}$ has received at least one LF vote for any class, i.e., if not all components of the corresponding $\boldsymbol{v}$ are -1, $\boldsymbol{d}$ belongs to $\mathcal{D}_\mathrm{l}$, otherwise $\mathcal{D}_\mathrm{u}$. In addition, let $\mathcal{V}_\mathrm{l}$ be the set consisting of $\boldsymbol{v}$ corresponding to $\boldsymbol{d}$ belonging to $\mathcal{D}_\mathrm{l}$. For the following steps, let $E \in \mathbb{R}^{\mathrm{n}(\mathcal{Y}) \times \mathrm{n}(\mathcal{L})}$ be the expanded vote matrix whose (c, l)-th component e_{cl} is 1 if l-th LF votes for class c, otherwise 0, and $\boldsymbol{b} \in \mathbb{R}^{\mathrm{n}(\mathcal{L})}$ be the binarized vote vector whose l-th component is 1 if the l-th LF votes for any class, otherwise 0.

The second step is training and validating soft attention BNN. Figure 2 shows the BNN structure. It receives $\mathcal{D}_\mathrm{l}$ and $\mathcal{V}_\mathrm{l}$ for training. Let $\mathcal{W}{=}\{W_\mathrm{c1}, W_\mathrm{c2}, W_\mathrm{a1}, W_\mathrm{a2}\}$ denote the set of weight parameter matrices, $\{\boldsymbol{m}_\mathrm{in}, \boldsymbol{m}_\mathrm{c}, \boldsymbol{m}_\mathrm{a}\}$ denote the cMCdropout masks, and $\tilde{\boldsymbol{d}} = \boldsymbol{d} {\odot} \boldsymbol{m}_\mathrm{in}$, where $\odot$ is Hadamard product, i.e., element-wise multiplication. The classifier estimation $p(\boldsymbol{y}_\mathrm{c})$ is calculated by a two-layer feedforward neural network as

$$p(\boldsymbol{y}_\mathrm{c}){=}\mathrm{softmax}(W_\mathrm{c2}^\mathsf{T}(\mathrm{ReLU}(W_\mathrm{c1}^\mathsf{T}\tilde{\boldsymbol{d}}){\odot}\boldsymbol{m}_\mathrm{c})). \quad (1)$$

Let $\tilde{\boldsymbol{v}}$ be a concatenation of $\tilde{\boldsymbol{d}}$ and $\boldsymbol{v}$ corresponding to $\boldsymbol{d}$. To reduce the label noise of WS, the LF reliability $\boldsymbol{a}{\in}\mathbb{R}^{\mathrm{n}(\mathcal{L})}$ whose l-th component a_l represents the reliability of the l-th LF is calculated by the other two-layer feedforward neural network as

$$\boldsymbol{a} = \mathrm{softmax}(W_\mathrm{a2}^\mathsf{T}(\tanh(W_\mathrm{a1}^\mathsf{T}\tilde{\boldsymbol{v}}) \odot \boldsymbol{m}_\mathrm{a})). \quad (2)$$

Then, the c-th component of attention estimation $p(\boldsymbol{y}_\mathrm{a})$ is calculated by the LF voting results weighted by each LF reliability as

$$p(y_\mathrm{a}{=}c) = \frac{\exp\left(\sum_l e_{cl} a_l\right)}{\sum_{c'} \exp\left(\sum_l e_{c'l} a_l\right)}. \quad (3)$$

To obtain a pseudo label $\hat{y}$, the averaged LF reliability $\bar{\boldsymbol{a}}$ is calculated through $\mathcal{D}_\mathrm{l}$ as

$$\bar{\boldsymbol{a}} = \frac{1}{\mathrm{n}(\mathcal{D}_\mathrm{l})} \sum_{\boldsymbol{d} \in \mathcal{D}_\mathrm{l}} (\boldsymbol{b} \odot \boldsymbol{a}). \quad (4)$$

Then, $\hat{y}$ is decided by weighted majority voting as

$$\hat{y} = \operatorname*{argmax}_c \sum_l e_{cl} \bar{a}_l. \quad (5)$$

The training is performed with $\hat{y}$ as pseudo ground truth and the loss defined as the weighted sum of negative log likelihoods (NLL) from $(\hat{y}, p(\boldsymbol{y}_\mathrm{c}))$ and $(\hat{y}, p(\boldsymbol{y}_\mathrm{a}))$ for co-training the classifier and the WS denoiser. Let $\hat{\mathcal{W}}$ denote the optimized weight parameters through the training. Thanks to the pseudo labeling technique, validation can be performed with no label leakage. A part of the unlabeled dataset is separated as a validation dataset in advance. The validation is performed with the pseudo ground truth obtained by the denoised WS as in training.

The third step is estimating posterior probability of $\mathcal{D}_\mathrm{u}$ on the optimized classifier with $\hat{\mathcal{W}}$. The output posterior probability distribution of $\boldsymbol{d}{\in}\mathcal{D}_\mathrm{u}$ considering the uncertainty is estimated by repeated estimation with a cMCdropout as $p(\boldsymbol{y}|\boldsymbol{d}, \hat{\mathcal{W}})$.

The fourth step is extracting semantic clusters by density clustering. This step also receives $\mathcal{D}_\mathrm{u}$. Let $\mathcal{C}_k \subset \mathcal{D}_\mathrm{u}$ denote the set of data samples belonging to the k-th cluster. Note that the maximum number of k changes depending on the clustering result. This step enables the following sampling to suggest data samples with semantic consistency to elicit probable labeling rules for humans to design LFs with lighter workload because the data in the identical cluster are more likely to have common semantic features.

The fifth step is sampling. Let $\mathcal{D}_{\mathrm{s}}$ and $q(\mathcal{D}_{\mathrm{s}})$ denote a small number of data samples selected from $\mathcal{D}_{\mathrm{u}}$ and its acquisition function calculated by batchBALD, respectively. Calculation of batchBALD involves using $p(\boldsymbol{y}|\boldsymbol{d}, \hat{\mathcal{W}})$ estimated in the third step. The difference from the original batchBALD is the clustering constraint added to make the sampling human-friendly. The batch is formed by only data samples belonging to an identical semantic cluster to ensure semantic consistency in the batch. Thus, in the implementation, batchBALD is applied for each $\mathcal{C}_k$ extracted in the fourth step. The candidate data samples $\tilde{\mathcal{D}}_{\mathrm{s}k}$ for each $\mathcal{C}_k$ are extracted as

$$\tilde{\mathcal{D}}_{\mathrm{s}k} = \underset{\mathcal{D}_{\mathrm{s}} \subset \mathcal{C}_k}{\operatorname{argmax}}\, q(\mathcal{D}_{\mathrm{s}}). \tag{6}$$

Among these candidates, the one with the largest batchBALD score is selected for the set of data samples $\tilde{\mathcal{D}}_{\mathrm{s}}$ to suggest for humans as

$$\tilde{\mathcal{D}}_{\mathrm{s}} = \underset{\tilde{\mathcal{D}}_{\mathrm{s}k}}{\operatorname{argmax}}\, q(\tilde{\mathcal{D}}_{\mathrm{s}k}). \tag{7}$$

BatchBALD with a clustering constraint is the key to reducing the number of LFs and the workload of LF designing simultaneously. Ratner et al. (2017) remarks that human defined LF tends to have low accuracy and large coverage. Trying to prevent it leads to make the coverage too small. The proposed framework can optimize the accuracy and the coverage by suggesting multiple data samples with common semantic features.

The last step is implementing a new LF. This is the only step performed by humans. Humans implement a new LF by referring to $\tilde{\mathcal{D}}_{\mathrm{s}}$ and add it to $\mathcal{L}$. The $\tilde{\mathcal{D}}_{\mathrm{s}}$ should have both semantic consistency and a large utility in classification performance when correctly labeled by the new LF. This process is an analogy for programming-by-examples in intuition-based human programming since the entity of LF is a simple program (Devlin et al., 2017). Thus, humans can implement an effective and probable LF with light workload. Thereafter, the loop consisting of these six steps is repeated the designated number of times while updating each variable associated with the addition of the new LF.

Table 1: Major parameter settings

The dimension of each hidden layer	128
learning rate	0.001
# of samples for cMCdropout	100
The maximum epoch for the training	1000
The patience of early stopping	100
The minimum cluster size	5
# of suggested samples in each loop	5

4 Experiment

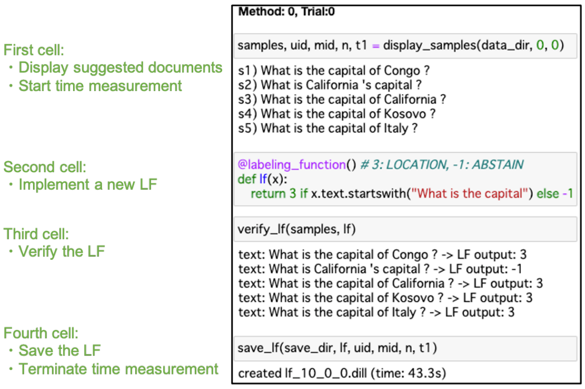
The object of the user studies with 16 consenting volunteers in two text classification tasks is to verify the effectiveness of the proposed method. The tasks are SMS spam detection (Almeida et al., 2011) and TREC-6 question classification (Li and Roth, 2002), which were also used in recent WS studies (Awasthi et al., 2020). These tasks do not require specialized domain knowledge except for basic skills in English reading comprehension and Python programming. The volunteers were non-native English speaking software engineers who met the above requirements and offered from multiple groups. They were likely to take on the labeling task and understand the task but did not know about WS. This condition is desirable to verify the effectiveness of the method with many subjects, even though it would be ideal to verify the method on tasks in which specialized domain experts work well, such as medical data or financial data. Note that automatic experiments without humans do not make sense because the objective of the proposed method is to enable humans to design probable LFs with light workload.

4.1 Settings

The number of classes and data samples for training, validation, and test are 2, 4504, 500, 500 in SMS and 6, 4906, 546, 500 in TREC-6, respectively. Note that no data include any ground truth labels in either training or validation. Table 1 shows the major parameter settings of the framework used in the user study. Hyper-parameters are not pre-optimized to the tasks considering the constraints of AL for WS from scratch in the wild. It is ideal to search for the best hyper-parameters with humans, but this is impractical in terms of human workload. Thus, it was also verified whether the hyper-parameters affect the superiority of the proposed method by recalculation using the implemented LFs. This setting is promising because the effectiveness of the implemented LFs is more important in this experimental design than the classification performance during the user study.

The implementation used sentence BERT (Reimers and Gurevych, 2019) provided by Hugging Face[1] for feature extraction, RAdam (Liu et al., 2020) for the optimizer, and HDBSCAN (Campello et al., 2013; McInnes and Healy, 2017) for the density clustering. The primary motivation for these choices is that they are robust to hyper-parameters and tolerant of overfitting in the state-of-the-art natural language processing (NLP) setting. Although it is common to use pre-trained BERT models for NLP tasks, the wider domains where the proposed method can be applied do not necessarily enable the use of prior knowledge. Thus, it was also verified whether the pre-training affects the superiority of the proposed method by recalculation using the features extracted by doc2vec (PV-DM) (Le and Mikolov, 2014), which is a representative method without pre-training.

To quantitatively verify the effectiveness of the method, the proposed method (Proposed) was compared to two other methods: batch extended Saisho et al. (2021) (Active) as the state-of-the-art AL for WS from scratch method and random sampling with a clustering constraint (Semantic) for reference. Note that the comparison with the other three types in the related work is meaningless because the prebuilt resources used in those types are the very targets or even more sophisticated ones expected to be implemented by subjects in the user study. "Semantic" randomly extracts data samples from the randomly extracted identical cluster. Simple random sampling (without the clustering constraint) should also be performed for an ablation study. However, it was excluded from the verification to prevent increasing the burden on subjects unnecessarily in the user study, considering that many existing studies have already shown AL is superior to random sampling. Similarly, to clarify in advance that the work time for each subject would fit into the daily working hours, the number of loops was set to 15, i.e., the total number of implemented LFs in each task is 45. The main evaluation metrics are the macro F-measure for classification performance, the time required to implement an LF, the accuracy of an LF voting, and the coverage of an LF voting.



Method: 0, Trial:0

```
samples, uid, mid, n, t1 = display_samples(data_dir, 0, 0)
```

```
s1) What is the capital of Congo ?
s2) What is California 's capital ?
s3) What is the capital of California ?
s4) What is the capital of Kosovo ?
s5) What is the capital of Italy ?
```

```
@labeling_function() # 3: LOCATION, -1: ABSTAIN
def lf(x):
  return 3 if x.text.startswith("What is the capital") else -1
```

```
verify_lf(samples, lf)
```

```
text: What is the capital of Congo ? -> LF output: 3
text: What is California 's capital ? -> LF output: -1
text: What is the capital of California ? -> LF output: 3
text: What is the capital of Kosovo ? -> LF output: 3
text: What is the capital of Italy ? -> LF output: 3
```

```
save_lf(save_dir, lf, uid, mid, n, t1)
```

```
created lf_10_0_0.dill (time: 43.3s)
```

Figure 3: Four filled cells to implement a new LF in each loop: humans can display the suggested samples and then implement, verify, and save a new LF.

Subjects were to work on the Jupyter notebooks (Kluyver et al., 2016) provided for each task as shown in Figure 3. When the first cell is run, the five data samples suggested by a method are displayed, and the time measurement starts. Subjects then implement an LF in a free format using Snorkel Python library (Ratner et al., 2017; Ratner et al., 2019) in the second cell, check its output by running the third cell, and repeat if they need to modify it. Finally, the fourth cell is run to send the new LF to a computer, and the time measurement is terminated. The computer then trains the classifier and selects the next suggested samples. Then, subjects can proceed to the next loop.

To evaluate the three methods equally, the user study devised two procedures. First, the subjects were told to perform the experiment procedure in the IMDB sentiment analysis (Maas et al., 2011) beforehand without performance measurement. This helps to reduce the effect on the results due to unfamiliarity with WS and the experimental procedure. Second, subjects were told to perform all three methods in parallel in a random order, without knowing the

[1] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v1

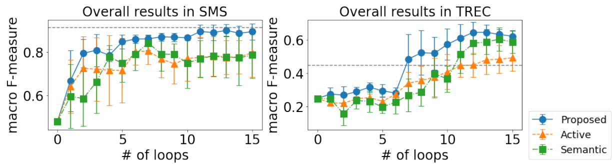


Figure 4: F-measures of each loop on each method: Proposed outperformed the other methods in terms of classification performance in all plots for both tasks.

Table 2: The mean time [sec.] and its relative value required to implement an LF: Proposed and Semantic outperformed Active thanks to the clustering constraint.

	SMS	TREC
Proposed	$\mathbf{156 \pm 143}$	$\mathbf{192 \pm 125}$
	$\mathbf{0.63 \pm 0.23}$	$\mathbf{0.75 \pm 0.05}$
Active	221 ± 166	258 ± 179
	1	1
Semantic	168 ± 137	197 ± 157
	0.78 ± 0.28	$\mathbf{0.75 \pm 0.10}$

differences between methods. This is also to prevent unfairness between methods due to mastery of WS and leakage of knowledge about the dataset, which may occur if the methods are conducted in series.

4.2 Results

Figure 4 shows the results of classification performance. Proposed outperformed the other methods in all plots in both tasks. These results quantitatively indicate that the proposed method enables humans to implement LFs with high utility. Slightly surprisingly, Semantic performed so well that its superiority to Active was intersected by plots. This viewpoint needs to be taken together with the performance of each LF, so it will be described in later paragraphs. In addition, the gray dashed lines show the results when applying all 73 LFs for SMS and 68 LFs for TREC from a recent WS study (Awasthi et al., 2020) to the classification model. The performance of Proposed is close to that for SMS and better than that for TREC. Note that the difference from the original results mainly comes from using some labeled data and a different evaluation metric in the original study. These results also indicate the effectiveness of the proposed method.

Table 2 shows the mean work times required to implement an LF. Their relative values on the basis of the mean of Active for each subject are also listed since there are large individual differences in the overall work time, Proposed reduced the work times by 37% in SMS and 25% in TREC compared with Active, respectively. These results quantitatively indicate that Proposed and Semantic are superior to Active in suggesting samples that elicit labeling rules for designing a new LF with a lighter workload. Thus, they show that the clustering constraint works well to make the framework human-friendly.

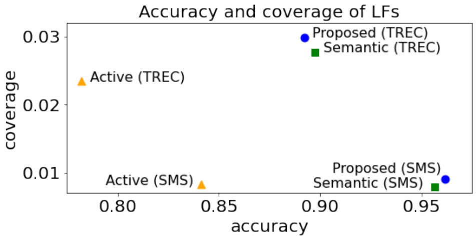
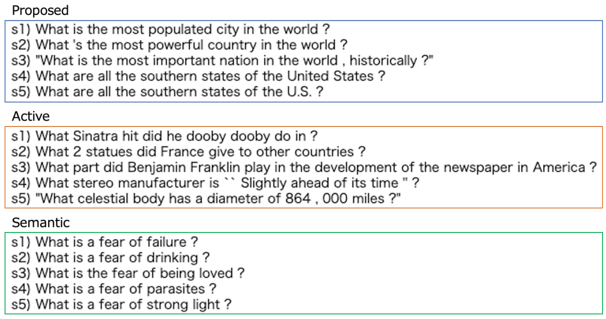


Figure 5: The mean accuracy and coverage of each LF: Proposed and Semantic outperformed Active in terms of helping humans to implement probable LFs.

Figure 5 shows the mean accuracy and mean coverage of each LF. Accuracy is the rate of votes that match the ground truth among the data samples in which LF votes. Coverage is the rate of data samples in the training dataset that were voted to any class by a single LF. In terms of coverage, there was no significant difference between the methods in SMS, so the superiority of the methods cannot be determined. On the other hand, there is a large difference in accuracy between Active and other methods, i.e., depending on whether the clustering constraint is added or not. This explains not only the reason Proposed is superior to other methods in the F-measure shown in Figure 4, but also the reason Semantic performs the same as or better than Active. These results indicate the importance of considering the accuracy of LF in a true human-in-the-loop situation in WS and the superiority of the proposed method to help humans design probable LFs.

For reference, Figure 6 shows the examples of data samples suggested at the same time by each method. Although it is difficult to explicitly visualize the difference in labeling utility by data samples, it should show the difference in the easiness of implementing a probable LF.

Figure 7 shows the results of hyper-parameter search by recalculations. The search spaces were $\{32, 64, 128, 256\}$ for the dimension of each hid-



Proposed

s1) What is the most populated city in the world ?
s2) What 's the most powerful country in the world ?
s3) "What is the most important nation in the world , historically ?"
s4) What are all the southern states of the United States ?
s5) What are all the southern states of the U.S. ?

Active

s1) What Sinatra hit did he dooby dooby do in ?
s2) What 2 statues did France give to other countries ?
s3) What part did Benjamin Franklin play in the development of the newspaper in America ?
s4) What stereo manufacturer is `` Slightly ahead of its time '' ?
s5) "What celestial body has a diameter of 864 , 000 miles ?"

Semantic

s1) What is a fear of failure ?
s2) What is a fear of drinking ?
s3) What is the fear of being loved ?
s4) What is a fear of parasites ?
s5) What is a fear of strong light ?

Figure 6: Examples of data samples suggested at the same time by each method in TREC task

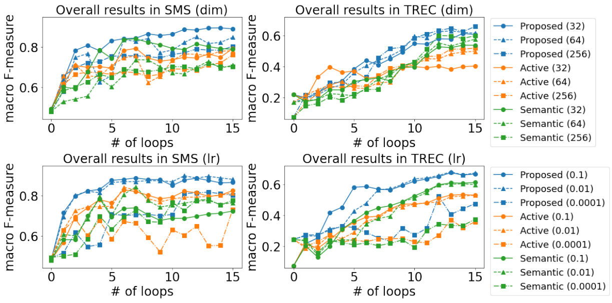


Figure 7: F-measures of each loop on each method with hyper-parameter search: the superiority of Proposed is not significantly affected under each condition although the superiority of a few plots is reversed. (Error bars are omitted in favour of visibility.)

den layer (dim) and $\{0.0001, 0.001, 0.01, 0.1\}$ for the learning rate (lr). When one hyper-parameter was searched for, the other was fixed to its default.

Lastly, Figure 8 shows the results obtained by recalculations with feature vectors extracted by doc2vec (PV-DM). In both recalculations, the overall trends do not differ significantly from the experimental results shown in Figure 4 under each condition although the superiority of a few plots is reversed and the overall performance changes. Thus, these results indicate that the superiority of the proposed method does not depend on hyper-parameters or the pre-trained embedding.

From the results of the user studies, the proposed method enables humans to implement more accurate LFs that can attach labels with greater utility for classification performance in a shorter work time.

4.3 Limitations and Future outlooks

In this section, three future outlooks are presented in accordance with the limitations of this paper. First,

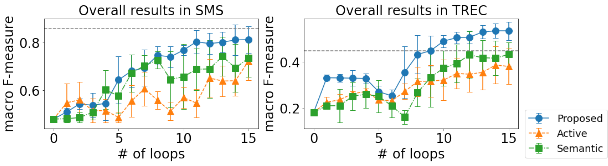
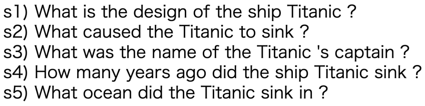


Figure 8: F-measures of each loop on each method without pre-trained embedding: the superiority of Proposed is not significantly affected although the superiority of a few plots is slightly reversed.



s1) What is the design of the ship Titanic ?
s2) What caused the Titanic to sink ?
s3) What was the name of the Titanic 's captain ?
s4) How many years ago did the ship Titanic sink ?
s5) What ocean did the Titanic sink in ?

Figure 9: Unsuitable example suggested by Proposed: ”Titanic” is a common feature among these samples, but it is not a key in the classification. The samples actually belong to several classes.

the common features extracted by semantic clustering may not always be effective for classification tasks. In a few cases, the semantic clustering constraint did not work well and suggested samples unsuitable for the classification task as shown in Figure 9. This indicates another support framework is needed for learning the humans’ viewpoint from the implemented LFs to make the semantic cluster constraint more directly related to LF implementation support. For example, a supervised attention approach is discussed also in a recent study (Sen et al., 2020). Incorporating explainable artificial intelligence (XAI) such as SHAP (Lundberg and Lee, 2017) should also be effective in eliciting more explicit rules for human-friendly support.

Second, errors can be reduced but not eliminated and will accumulate further. The classification performance was rarely degraded by adding an LF. This is because negative effects of noisy LFs are much larger than those of noisy labels in standard labeling. In AL, some solutions for noisy labels have been devised (Bouguelia et al., 2015; Bouguelia et al., 2018; Lin et al., 2016). The framework that extends these methods to noisy LFs should be very practical.

Lastly, some individual human differences in knowledge and skills cannot be adequately supported, and the experiment was performed by only software engineers. It is true that the LF implementation task depends more on individual skills than the traditional direct labeling task. Thus, methods to help LF implementation without programming such as Hancock et al. (2018) should be effective in compensating for people's lack of skills. Beyond that, the development of tools to integrate the proposed method with autoML and no-code development tools will lead to the democratization of NLP as well as AI. From another point of view, task distribution including both LF implementation and direct labeling is also an alternative research direction when multiple humans perform a single task, especially as in crowdsourcing. Such frameworks can provide an environment where anyone can work in accordance with their skills.

5 Conclusion

This paper proposed a new method of practical active learning (AL) for weak supervision (WS) in the wild from scratch and demonstrated its superiority in helping users to design and implement labeling functions (LFs) in terms of classification performance, required work time, and LF accuracy in user studies with a real interactive system and no label leakage. Its superiority is achieved by Bayesian AL with a semantic clustering constraint for enhancing labeling utility, reducing human workload, and enhancing the effectiveness of each LF.. For future work, the model will be verified by other complicated tasks and larger-scale user studies, in addition to the future outlook.

References

T. Almeida, J. Hidalgo, and A. Yamakami. 2011. Contributions to the study of sms spam filtering: New collection and results. In *Proc. of DocEng*, pages 259–262.

P. Atighehchian, F. Charron, and A. Lacoste. 2020. Bayesian active learning for production, a systematic study and a reusable library. In *Proc. of ICML workshops*.

A. Awasthi, S. Ghosh, R. Goyal, and S. Sarawagi. 2020. Learning from rules generalizing labeled exemplars. In *Proc. of ICLR*.

S. Bach, B. He, A. Ratner, and C. Ré. 2017. Learning the structure of generative models without labeled data. In *Proc. of ICML*, pages 273–282.

S. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, R. Kuchhal, C. Ré, and R. Malkin. 2019. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proc. of SIGMOD*, pages 362–375.

S. Biegel, R. Khatib, L. Oliveira, M. Baak, and N. Aben. 2021. Active weasul: Improving weak supervision with active learning. In *Proc. of ICLR workshops*.

B. Boecking, W. Neiswanger, E. Xing, and A. Dubrawski. 2021. Interactive weak supervision: Learning useful heuristics for data labeling. In *Proc. of ICLR*.

M. Bouguelia, Y. Belaïd, and A. Belaïd. 2015. Identifying and mitigating labelling errors in active learning. In *Proc. of ICPRAM*, pages 35–51.

M. Bouguelia, S. Nowaczyk, K. Santosh, and A. Verikas. 2018. Agreeing to disagree: active learning with noisy labels without crowdsourcing. *Int. J. Mach. Learn. Cybern.*, 9:1307–1319.

R. Campello, D. Moulavi, and J. Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Proc. of KDD*, pages 160–172.

W. Chen, Y. Liu, Z. Kira, Y. Wang, and J. Huang. 2019. A closer look at few-shot classification. In *Proc. of ICLR*.

N. Das, S. Chaba, R. Wu, S. Gandhi, D. Horng Chau, and X. Chu. 2020. Goggles: Automatic image labeling with affinity coding. In *Proc. of SIGMOD*, pages 1717–1732.

S. Dasgupta. 2005. Coarse sample complexity bounds for active learning. In *Proc. of NeuIPS*, pages 235–242.

J. Devlin, J. Uesato, S. Bhupatiraju, R. Singh, A. Mohamed, and P. Kohli. 2017. Robustfill: Neural program learning under noisy i/o. In *Proc. of ICML*, pages 990–998.

J. Fries, P. Varma, V. Chen, K. Xiao, H. Tejeda, P. Saha, J. Dunnmon, H. Chubb, S. Maskatia, M. Fiterau, S. Delp, E. Ashley, C. Re, and J. Priest. 2019. Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature Communications*, 10(3111).

Y. Gal and Z. Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. of ICML*, volume 48, pages 1050–1059.

J. Gonsior, M. Thiele, and W. Lehner. 2020. Weakal: Combining active learning and weak supervision. In *Proc. of DS*, pages 34–49.

B. Hancock, P. Varma, S. Wang, M. Bringmann, P. Liang, and C. Ré. 2018. Training classifiers with natural language explanations. In *Proc. of ACL*, pages 1884–1895.

A. Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv:1112.5745*.

P. Ipeirotis, F. Provost, V. Sheng, and J. Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Min. Knowl. Discov.*, 28(2):402–441.

D. Kartchner, W. Ren, D. An, C. Zhang, and C. Mitchell. 2020. Regal: Rule-generative active learning for model-in-the-loop weak supervision. In *Proc. of NeurIPS workshops*.

A. Kirsch, J. van Amersfoort, and Y. Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proc. of NeurIPS*, volume 32.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and Jupyter development team. 2016. Jupyter notebooks - a publishing format for reproducible computational workflows. In *Proc. of ElPub*, pages 87–90.

Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML*, pages 1188–1196.

P. Lertvittayakumjorn, L. Specia, and F. Toni. 2020. FIND: Human-in-the-Loop Debugging Deep Text Classifiers. In *Proc. of EMNLP*, pages 332–348.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proc. of COLING*, pages 1–7.

G. Li. 2017. Human-in-the-loop data integration. *Proc. of VLDB Endow.*, 10(12):2006–2017.

C. Lin, Mausam, and D. Weld. 2016. Re-active learning: Active learning with relabeling. In *Proc. of AAAI*, pages 1845–1852.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proc. of ICLR*.

S. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Proc. of NeurIPS*.

A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. 2011. Learning word vectors for sentiment analysis. In *Proc. of ACL*, pages 142–150.

L. McInnes and J. Healy. 2017. Accelerated hierarchical density based clustering. In *Proc. of ICDM workshops*, pages 33–42.

M. Nashaat, A. Ghosh, J. Miller, S. Quader, C. Marston, and J. Puget. 2018. Hybridization of active learning and data programming for labeling large industrial datasets. In *Proc. of IEEE BigData*, pages 46–55.

M. Nashaat, A. Ghosh, J. Miller, and S. Quader. 2020. Asterisk: Generating large training datasets with automatic active supervision. *Trans. Data Sci.*, 1(2).

K. Qian, P. Raman, Y. Li, and L. Popa. 2020. Learning structured representations of entity names using ActiveLearning and weak supervision. In *Proc. of the EMNLP*, pages 6376–6383.

M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. In *Proc. of NeurIPS*, pages 3347–3357.

A. Ratner, C. Sa, S. Wu, D. Selsam, and C. Ré. 2016. Data programming: Creating large training sets, quickly. In *Proc. of NeurIPS*, pages 3567–3575.

A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. 2019. Training complex models with multi-task weak supervision. In *Proc. of AAAI*, pages 4763–4771.

N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of EMNLP*, pages 3982–3992.

W. Ren, Y. Li, H. Su, D. Kartchner, C. Mitchell, and C. Zhang. 2020. Denoising multi-source weak supervision for neural text classification. In *Proc. of EMNLP*, pages 3739–3754.

O. Saisho, T. Ohguro, J. Sun, H. Imamura, S. Takeuchi, and D. Yokozeki. 2021. Human knowledge based efficient interactive data annotation via active weakly supervised learning. In *Proc. of PerCom workshops*, pages 332–335.

C. Sen, T. Hartvigsen, B. Yin, X. Kong, and E. Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proc. of ACL*, pages 4596–4608.

B. Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

A. Siddhant and Z. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proc. of EMNLP*, pages 2904–2909.

P. Varma and C. Ré. 2018. Snuba: Automating weak supervision to label training data. *Proc. VLDB Endow.*, 12(3):223–236.

B. Wang, A. Ratner, S. Mussmann, and C. Re. 2019. Interactive programmatic labeling for weak supervision. In *Proc. of KDD workshop*.

J. Yuan, X. Hou, Y. Xiao, D. Cao, W. Guan, and L. Nie. 2019. Multi-criteria active deep learning for image classification. *Knowledge-Based Systems*, 172:86–94.

Z. Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

SMTCE: A Social Media Text Classification Evaluation Benchmark and BERTology Models for Vietnamese

Luan Thanh Nguyen[1,2], **Kiet Van Nguyen**[1,2], **Ngan Luu-Thuy Nguyen**[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{luannt, kietnv, ngannlt}@uit.edu.vn

Abstract

Text classification is a typical natural language processing or computational linguistics task with various interesting applications. As the number of users on social media platforms increases, data acceleration promotes emerging studies on **S**ocial **M**edia **T**ext **C**lassification (**SMTC**) or social media text mining on these valuable resources. In contrast to English, Vietnamese, one of the low-resource languages, is still not concentrated on and exploited thoroughly. Inspired by the success of the GLUE, we introduce the **S**ocial **M**edia **T**ext **C**lassification **E**valuation (**SMTCE**) benchmark, as a collection of datasets and models across a diverse set of SMTC tasks. With the proposed benchmark, we implement and analyze the effectiveness of a variety of multilingual BERT-based models (mBERT, XLM-R, and DistilmBERT) and monolingual BERT-based models (PhoBERT, viBERT, vELECTRA, and viBERT4news) for tasks in the SMTCE benchmark. Monolingual models outperform multilingual models and achieve state-of-the-art results on all text classification tasks. It provides an objective assessment of multilingual and monolingual BERT-based models on the benchmarks, which will benefit future studies about BERTology in the Vietnamese language.

1 Introduction

With the rise of social media, researching user comments may allow us to understand their behavior and enhance the quality of cyberspace. Social text classification is one of the current popular NLP tasks that aim to solve that problem with an extensive study on comments users left on social media platforms. There are several different current tasks in social media text classification (SMTC) tasks in several domains, and however, it is designed to be independent of other inconsistent data domains.

Motivated by the popularity of the GLUE (Wang et al., 2018), we present a novel Social Media Text Classification Evaluation (SMTCE) benchmark with four social media text classification tasks in Vietnamese, including constructive speech detection (Nguyen et al., 2021), complaint comment dectection (Nguyen et al., 2022a), emotion recognition (Ho et al., 2020), and hate speech detection (Luu et al., 2021) tasks.

Natural Language Processing (NLP), a subset of artificial intelligence techniques, is rapidly evolving, with numerous notable accomplishments. The primary goal of assisting computers in understanding human language is the bridge for reducing the distance between humans and computers. The current trend in NLP is releasing language models that have been pre-trained on a considerable amount of data to perform various tasks. Hence, we do not need to code from scratch for every task and annotate millions of texts each time. This is the premise for transfer learning and applying the architectures transformers to various downstream tasks in NLP.

Bidirectional Encoder Representations from Transformers, known as BERT, have appeared and played an essential role in the current outstanding development of computational linguistics. BERT has become a typical baseline in NLP experiments. Several numerous kinds of research have been conducted to analyze the performance of BERTology models (Rogers et al., 2020). Following its release, multilingual BERT-based models emerged, achieving breakthrough performances in serving several languages. Moreover, monolingual language models are also concentrated, primarily focused on a single language, particularly a low-resource language like Vietnamese. In this study, we conduct experiments with multilingual and monolingual BERT-based language models on the SMTCE benchmark to see how they perform.

Our main contributions to this research are:

- We propose **SMTCE**, the **S**ocial **M**edia **T**ext **C**lassification **E**valuation benchmark for evaluating social media text classification or social media mining in Vietnamese.

- We implement various BERT-based models on the SMTCE benchmark. We then compare and contrast the characteristics and strengths of monolingual and multilingual BERT-based language models on computing Vietnamese text classification tasks.

- After achieving results of monolingual language models, which outperform multilingual models, we start to discuss the remaining problems that researchers on social media text mining have to face.

The remainder of this work is structured as follows: Section 2 includes related works to which we refer. Section 3 describes the tasks in the SMTCE benchmark. Section 4 overviews BERT-based language models available for Vietnamese NLP tasks. Section 5 shows the processes for implementing models and their results for each task in the SMTCE benchmark. Section 6 is the discussion of the remaining problems in this study. Section 7, the last section, is our conclusion for this research and future work.

2 Related Work

The introduction of BERT, Devlin et al. (2019) led to the explosion of the transformer language models. BERT has obtained state-of-the-art performances on a variety of NLP tasks upon launch, including nine GLUE tasks, SQuAD v1.0 and 2.0, and SWAG.

Shortly after releasing BERT, Devlin et al. (2019) then published multilingual BERT, which was capable of over 100 languages. Then, a slew of BERTology models started to emerge. CONNEAU and Lample (2019) released XLM, a cross-lingual language model achieving promising results on various NLP tasks. Following the introduction RoBERTa, they introduced a new pre-trained model, XLM-R (Conneau et al., 2020), which reached breakthrough results.

Aside from multilingual versions or BERT-based models, developing NLP in countries with different languages promotes researchers to build and improve monolingual models based on available BERT architectures for their languages. We have CamemBERT (Martin et al., 2020) for France, Chinese-BERT (Cui et al., 2020) for Chinese, or BERT-Japanese (Kikuta, 2019) for Japanese. One of the low-resource languages, Vietnamese, has monolingual BERT-based models that have been pre-trained on Vietnamese datasets, such as: PhoBERT (Nguyen and Tuan Nguyen, 2020), viBERT (Bui et al., 2020), vELECTRA (Bui et al., 2020), or viBERT4news[1].

In Vietnamese, To et al. (2021a) did research about investigating monolingual and multilingual BERT-based models for the Vietnamese summarization task. It is the first attempt to use datasets from other languages based on multilingual models to execute various existing pre-trained language models on the summarization task.

In this paper, we implement several monolingual and multilingual BERT-based pre-trained language models on the proposed SMTCE benchmark tasks. Ensuingly, we conduct an overview of these two types of models in Vietnamese SMTC tasks.

3 Social Media Text Classification Tasks

Technology is continuously changing, and social networks allow our users to interact and exchange information more easily. Because of this ease, many harmful and malicious comments of anonymous users aimed to attack individuals psychologically. Studies in this field are gaining attraction to help automatically classify comments as helpful, constructive, or harmful to block and hide them promptly. By giving suitable solutions depending on various situations, we hope to create positive and friendly cyberspace.

In this study, we propose SMTCE, a new benchmark concentrating on four social media text classification tasks, which covers various domains, data sizes, and challenges in social media tasks. Overview of the tasks in the SMTCE is shown in Table 1 with the statistics and descriptions of tasks in the SMTCE benchmark, including dataset name, the number of texts of each set, task target, inter-annotator agreement (IAA), baseline result, and data source.

3.1 Emotion Recognition Task

Ho et al. (2020) released a standard Vietnamese Social Media Emotion Corpus known as UIT-VSMEC (VSMEC) for solving the task of recognizing the emotion of Vietnamese comments on

[1] https://github.com/bino282/bert4news

Table 1: Statistics and descriptions of tasks in the SMTCE benchmark. All datasets used the macro-average F1-score to measure the performance of machine learning models.

Dataset	Train	Dev	Test	Task	IAA	Baseline Result (F1-macro %)	Data Source
Binary text classification							
ViCTSD	7,000	2,000	1,000	Constructive speech detection	0.59	78.59	News comments
ViOCD	4,387	548	549	Complaint comment detection	0.87	92.16	E-commerce feedback
Multi-class text classification							
VSMEC	5,548	686	693	Emotion recognition	0.80	59.74	Social network comments
ViHSD	24,048	2,672	6,680	Hate speech detection	0.52	62.69	Social network comments

social media. It is made up of 6,927 sentences that had been manually annotated with seven emotion labels, including anger, disgust, enjoyment, fear, sadness, surprise, and other. Figure 1 illustrates the number of sentences of each label in the dataset. As we can see, the highest number of sentences belongs to the label enjoyment with 1,965 sentences, and the lowest is the surprise label with 309 sentences.

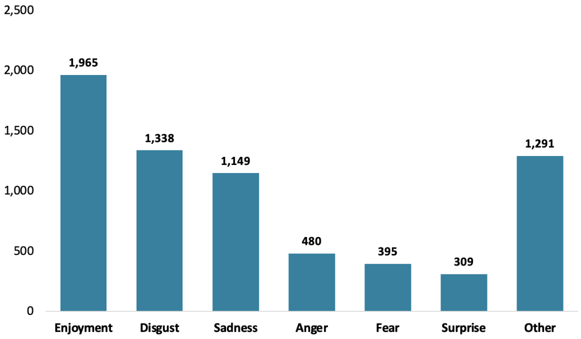


Figure 1: The number of sentences of each label in the VSMEC dataset.

3.2 Constructive Speech Detection Task

The dataset UIT-ViCTSD (ViCTSD: Vietnamese Constructive and Toxic Speech Detection) Nguyen et al. (2021) was built for dealing with the task of detecting constructive and toxic speech. This task aims to solve two issues: detecting constructive and toxic speech in Vietnamese social media comments. Each comment is labeled into two different tasks: identifying constructive and toxic comments. To define constructiveness, there are two labels: constructive and non-constructive. Furthermore, there are two labels for classifying toxic comments: toxic and non-toxic. Figure 2 below is the statistic of the number of each label in the dataset.

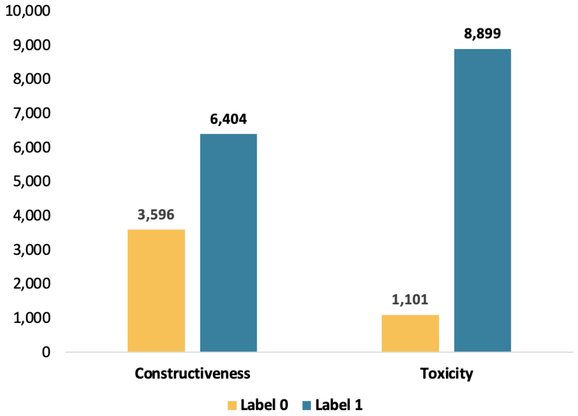


Figure 2: The statistic of the number of sentences of each label in the ViCTSD dataset.

The dataset contains 10,000 comments by ten domains of users crawled from the online discussion as VnExpress.net. The dataset serves to identify the constructiveness and toxicity of Vietnamese social media comments. The authors also evaluated the first version of this dataset with a proposed system with 78.59% and 59.40% F1-score for constructiveness and toxicity detection.

As depicted in Figure 2, we see that the dataset is very imbalanced in toxic speech detection tasks, so we only focus on the task of constructiveness detection in this study.

3.3 Hate Speech Detection Task

Luu et al. (2021) provided a dataset for the task of hate speech detection on Vietnamese social media comments named UIT-ViHSD (ViHSD). The dataset includes 30,000 comments labeled by annotators with three labels CLEAN, OFFENSIVE, and HATE. We describe the proportion of each label in the dataset in Figure 3. As shown in this illustration, this dataset is severely unbalanced (82.71% of CLEAN comments) with a low inter-annotator agreement of only 0.52, and it required several techniques in pre-processing data to deal with this imbalance.

Table 2: An overview of available BERTology languages model in Vietnamese.

		Data size	Vocab. size	Tokenization	Domain
Multilingual	mBERT (case uncased)	16GB	3.3B	Subword	Book+Wiki
	XLM-R (base)	2.5TB	250K	Subword	Common Crawl
	DistilmBERT	16GB	31K	Subword	Book+Wiki
Monolingual	PhoBERT	20GB	64K	Subword	News+Wiki
	viBERT	10GB	32K	Subword	News
	vELECTRA	60GB	32K	Subword	News
	viBERT4news	20GB	62K	Syllable	News

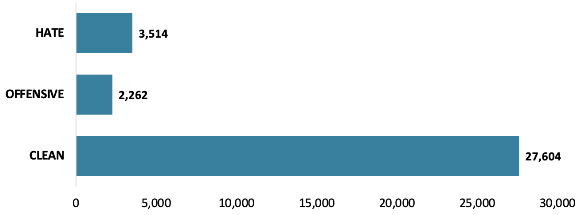


Figure 3: The statistic of the number of sentences of each label in the ViHSD dataset.

Hate speech has been a source of concern for social media users. The dataset aims to create a tool for identifying it in online communication interactions, censoring it to protect users from offensive content, and improving the environment of online forums.

3.4 Complaint Comment Detection Task

Nguyen et al. (2022a) researched customer complaints on e-commerce sites and released a novel open-domain dataset named UIT-ViOCD (ViOCD), a collection of 5,485 human-annotated comments on four domains. It was then evaluated using multiple approaches and achieving the best performance with an F1-score of 92.16% by the fine-tuned PhoBERT model. We can utilize this information to classify complaint comments from users on open-domain social media automatically. Table 3 below depicts the distribution of each label in sets.

Table 3: The distribution of each label in the train, valid, and test sets of the ViOCD dataset.

	Complaint	Non-complaint
Train set	2,292	2,095
Dev set	283	265
Test set	279	270
Total	**2,854**	**2,630**

Even though this dataset contains a small number of data points, the ratio between labels 0 and 1 is quite balanced. Hence, in this task, we do not need to use significant techniques to deal with data imbalance before starting the training session.

4 BERTology Language Models

Transformer-based models are currently the best-performance models in natural language processing or computational linguistics. The architecture of the transformer model consists of two parts: encoder and decoder. It differs from previous deep learning architectures that it pushes in all of the input data at once rather than progressively, thanks to the self-attention (Vaswani et al., 2017) mechanism. Following the premiere of Devlin et al. (2019), models based on BERT architectures are becoming increasingly popular. Challenging NLP tasks have been effectively solved using these models.

BERT (Bidirectional Encoder Representations from Transformers), the same architecture as the bi-directional recurrent neural network, uses bi-directional encoder representations instead of traditional techniques that only learn from left to right or right to left. A bi-directional architecture includes two networks, one in which the input is handled from start to end and another one from end to start. The outputs of two networks are then integrated to provide a single representation. As a result, BERT is better able to understand the relationship between words and provides better performances.

Besides, unlike other context-free word embedding models like word2vec or GloVe, which create a single word embedding representation for each word in its vocabulary, BERT needs to consider each context for a given the word in a phrase. As a consequence, homonyms in other sentences become different in each context.

In this study, we try to apply various BERT-based models to Vietnamese datasets for solving social media text classification tasks. With multilingual models, we implement multilingual BERT cased (mBERT cased), multilingual BERT uncased (mBERT uncased), XLM-RoBERTa (XLM-R), and Distil multilingual BERT (DistilmBERT). We also try the monolingual models, which are pre-trained in Vietnamese data, such as PhoBERT, viBERT, vELECTRA, and viBERT4news. Table 2 is an overview of multilingual and monolingual BERT-based language models available in Vietnamese. All pre-trained models we used in this research are downloaded from Hugging Face[2].

4.1 Multilingual Language Models

Language models are usually specifically designed and trained in English, most globally used language. Moreover, these models were then deeper trained and became multilingual to expand and serve NLP problems to support more languages globally. In this study, we deploy popular multilingual models as follows.

4.1.1 Multilingual BERT

Devlin et al. (2019) continues to develop and expand supported languages with multilingual BERT, including uncased and cased versions, after the launch of BERT. Multilingual BERT models, as opposed to their predecessors, are trained in various languages, including Vietnamese, and use masked language modeling (MLM). Each model includes a 12-layer, 768-hidden, 12-heads, and 110M parameters and supports 104 different languages. There are two multilingual BERT models: uncased[3] and cased[4] versions, which are both used for implementation in this study.

4.1.2 XLM-RoBERTa

XLM-RoBERTa (XLM-R) is a cross-lingual language model provided by Conneau et al. (2020) and trained in 100 different languages, including Vietnamese, on 2.5TB Clean CommonCrawl data[5]. It offers significant gains in downstream tasks such as classification, sequence labeling, and question answering over previously released multilingual models like mBERT or XLM. Under XLM, the language of the input ids cannot be correctly determined by the XLM-R language for the language of the input id to be understood by language tensors. We use the XLM-R base[6] in this research.

4.1.3 DistilmBERT

DistilmBERT[7], which is published by Sanh et al. (2020), is a different BERT version, in which its features improve the original version (faster, cheaper and lighter). This architecture requires less than BERT pre-training. In addition, it is more efficient compared to BERT, as the BERT model could be reduced by 40 while maintaining 97 of its language understanding capabilities and 60 more quickly. Hence, DistilmBERT has been known as a BERT version of reducing the number of parameters.

4.2 Monolingual Language Models

In addition to developing multilingual language models, researchers in specific languages are interested in monolingual models. Monolingual models are frequently built using the BERT architecture and pre-trained on datasets in a single language. Furthermore, because these models are trained on a large amount of data in a language, they frequently achieve great performances on NLP tasks for the languages themselves. In this study, we use several existing monolingual models that have been introduced for solving Vietnamese tasks.

4.2.1 PhoBERT

Nguyen and Tuan Nguyen (2020) first presented PhoBERT models, which are pre-trained models for Vietnamese NLP. They are state-of-the-art Vietnamese language models. To handle tasks in Vietnamese, they have trained the first large-scale monolingual BERT-based with two versions as *base* and *large* with a 20GB word-level Vietnamese dataset combining two datasets: Vietnamese Wikipedia[8] (1GB) and modified dataset from a Vietnamese news dataset[9] (19GB). The chosen version we use to implement in this study is PhoBERT base[10].

[2] https://huggingface.co/

[3] https://huggingface.co/bert-base-multilingual-uncased

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://github.com/facebookresearch/cc_net

[6] https://huggingface.co/xlm-roberta-base

[7] https://huggingface.co/distilbert-base-multilingual-cased

[8] https://vi.wikipedia.org/wiki/

[9] https://github.com/binhvq/news-corpus

[10] https://huggingface.co/vinai/phobert-base

Table 4: The techniques for pre-processing data we employed in experiments.

No.	Pre-procesing technique	Dataset			
		VSMEC	ViCTSD	ViHSD	ViOCD
1	Removing numbers	✓		✓	✓
2	Removing punctations		✓	✓	✓
3	Removing emojis, emoticons		✓	✓	✓
4	Converting emojis, emoticons into texts	✓			
5	Tokenizing words	✓	✓	✓	✓

4.2.2 viBERT

ViBERT[11], a pre-trained language model for Vietnamese, which is based on BERT architecture and introduced by Bui et al. (2020). The architecture of viBERT is similar to that of mBERT, and it has been pre-trained on large corpora of 10GB of uncompressed Vietnamese text. Nonetheless, there is a distinction between this model and mBERT. They chose to exclude insufficient vocab because the mBERT vocab still contains languages apart from Vietnamese.

4.2.3 vELECTRA

ELECTRA, first introduced by Clark et al. (2020), is a novel pre-training architecture that uses replaced token detection (RTD) rather than language modeling (LM) or masked language modeling (MLM), as is popular in existing language models.

Bui et al. (2020) also released vELECTRA, another pre-trained model for Vietnamese, along with viBERT. They used a dataset with almost 60GB of words to pre-train vELECTRA[12].

4.2.4 viBERT4news

NlpHUST published viBERT4news[13], a Vietnamese version of BERT trained on more than 20 GB of news datasets. ViBERT4news demonstrated its strength on Vietnamese NLP tasks, including sentiment analysis using comments of the AIViVN dataset, after launch and testing, with an F1 score of 0.90268 on public leaderboards (while the winner of the shared-task score is 0.90087).

5 Experiments and Results

In this section, we carry out experiments using monolingual and multilingual BERT-based models on Vietnamese benchmark datasets.

5.1 Pre-processing Techniques

We implement our data preprocessing strategies based on the task as well as the characteristics of the dataset. Because each dataset is distinct in terms of vocabulary, origin, and content, we use different preprocessing approaches appropriate for each dataset before feeding data into the models. Table 4 contains a list of techniques that we implement for each dataset.

We almost remove numbers in all tasks except the constructive speech detection task because of Nguyen et al. (2021) recommendation. For the VSMEC dataset, we do not remove punctuations, emojis, and emoticons because, in several cases, users often use them in their comments to express emotions, which has a great effect on the performance of the model in predicting emotion labels. Therefore, we keep the punctations, and we then apply the study of Nguyen and Van Nguyen (2020) in data preprocessing, which is transforming emojis and emoticons into Vietnamese text and then obtain a higher performance with the F1-score of 64.40% (higher than the baseline of the author of the dataset 4.66%). Because emojis and emoticons are also essential elements influencing the emotions expressed in comments, deleting them causes a harmful effect on the emotion categorization task.

There are several preprocessing methods with different transfer learning-based models as a recommendation by its authors. For PhoBERT, we employ VnCoreNLP[14] and FAIRSeq[15] for preprocessing data and tokenizing words before applying them to the dataset.

5.2 Experiment Settings

We record and select the most parameters for each task after multiple experiment and show them in Table 6. Additionally, we keep the max sequence

[11] https://huggingface.co/FPTAI/vibert-base-cased

[12] https://huggingface.co/FPTAI/velectra-base-discriminator-cased

[13] https://github.com/bino282/bert4news

[14] https://github.com/vncorenlp/VnCoreNLP

[15] https://github.com/pytorch/fairseq

Table 5: Experimental Results of multilingual versus monolingual models on Vietnamese social media datasets (macro-averaged F1-score (%)).

Model		VSMEC	ViCTSD	ViOCD	ViHSD
Baseline		59.74 (Ho et al., 2020)	78.59 (Nguyen et al., 2021)	92.164 (Nguyen et al., 2022a)	62.69 (Luu et al., 2021)
Multilingual	*mBERT (cased)*	54.59	80.42	91.61	64.20
	mBERT (uncased)	53.14	78.97	92.52	62.76
	XLM-R	62.24	80.51	94.35	63.68
	DistilmBERT	53.83	81.69	90.50	62.50
Monolingual	*PhoBERT*	**65.44**	83.55	94.71	66.07
	viBERT	60.68	81.27	94.53	65.06
	vELECTRA	61.29	80.24	**95.26**	65.97
	viBERT4news	64.65	**84.15**	94.72	**66.43**

length as used in the baseline of each task because they are optimized settings for it. Other parameters are remained and not customized.

Table 6: The parameters selected for each task in the SMTCE benchmark. [1] and [2] represent two parameters batch_size and epochs, respectively.

	VSMEC		ViOCD		ViHSD		ViCTSD	
	[1]	*[2]*	*[1]*	*[2]*	*[1]*	*[2]*	*[1]*	*[2]*
mBERT (cased)	16	4	16	4	16	4	16	4
mBERT (uncased)	16	4	16	4	16	4	16	4
XLM-R	8	4	16	4	16	4	16	2
DistilmBERT	16	4	16	4	16	4	16	4
PhoBERT	8	2	8	2	16	2	16	2
viBERT	16	4	16	4	16	4	16	4
vELECTRA	16	4	16	4	16	4	16	4
viBERT4news	8	2	16	2	16	2	16	2

5.3 Evaluation Metric

In the text classification task, we have different metrics suitable for specific datasets and problems. Because most datasets in this study are imbalanced and according to the choice of dataset authors, we choose the macro-average F1 score to evaluate the performances of models on the datasets.

To compute the macro-average F1 score, we first calculate the F1 score per class in the dataset by the formula (1).

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

After achieving the F1 scores of all classes, we compute the macro-average F1 score by calculating the average F1 score as shown in formula (2).

$$\text{Macro F1 score} = \frac{sum(\text{F1 scores})}{\text{Number of classes}} \quad (2)$$

Because the macro f1 score gives importance equally to each class, it means the macro f1 algorithm may still produce objective findings on unbalanced datasets since a majority class will participate equally alongside the minority. That is the reason why almost imbalanced dataset authors choose the macro-average f1 score as the primary metric to represent actual model performance despite skewed class sizes.

5.4 Experimental Results

We begin implementation after selecting appropriate parameters for each model corresponding to each task, and the results obtained by models are presented in Table 5.

Our experiments outperform the original results of authors on each task in the majority of cases. As a result, we obtain the following outcomes: 65.44% for VSMEC sentiment classification task (higher than 5.7%) with PhoBERT model; 84.15% for ViCTSD identifying constructiveness task (higher than 5.56%) by viBERT4news model; 95.26% for ViOCD classifying complaint comments task (higher than 3.1%) by vELECTRA model; 66.43% for ViHSD hate speech detection task (higher than 3.74%) by viBERT4news model. The results on the ViOCD set are significantly higher than on other datasets because its data samples are processed cleanly, and the higher inter-annotator agreement during data labeling is noticeable when building the dataset.

5.5 Result Analysis

After achieving the results, we see that the outcomes of monolingual models are all greater than the baseline result of the author for each dataset. In terms of performance, specific techniques, like XLM-R with VSMEC; all methods with ViCTSD; mBERT cased and uncased, and XLM-R with ViHSD; and mBERT uncased and XLM-R with

ViOCD, perform just above the baseline.

We also discover that the models that produce the most remarkable outcomes in this study for all tasks are monolingual. It demonstrates that monolingual language models outperform multilingual models when dealing with non-English and low-resource languages such as Vietnamese.

When we implement techniques on the VSMEC dataset, the best model is PhoBERT, which has an F1-score of 65.44%, more significant than the result of the author of 4.66%. Huynh et al. (2020) also published a method that gains an F1-score of 65.79% (higher than ours 0.35%). However, the approach used in that study is an ensemble classifier that combines multiple Neural Network Models such as CNN, LSTM, and mBERT. This ensemble approach must operate with a majority voting ensemble before providing the final output, which is why it takes longer to execute than a single model.

6 Discussions on Vietnamese Social Media Text Mining

6.1 Monolingual versus Multilingual: Which is the better in Vietnamese social media text classification on BERTology?

After implementing multiple multilingual and monolingual BERT-based language models for each task in the SCTE benchmark, we discovered that monolingual gets better results in most cases. Furthermore, the results of all current monolingual models outperform the baseline of the author of the dataset. Meanwhile, multilingual models only outperform baseline in a few tasks and mainly stand out with XLM-R.

Multilingual models do not now provide superior results because these models, except XLM-R, are primarily pre-trained on the Vietnamese Wikipedia corpus. This corpus is limited, with only 1GB of uncompressed data, and the material on Wikipedia is not representative of common language use. Moreover, unlike in English or other widely spoken languages, the space in Vietnamese is solely used to separate syllables, not words. Meanwhile, multilingual BERT-based models are now unaware of this. Monolingual models are pre-trained on larger and higher-quality Vietnamese datasets, such as viBERT4news on AIViVN's comments dataset or PhoBERT on a 20GB word-level Vietnamese corpus (Nguyen and Tuan Nguyen, 2020). As a result, monolingual language models outperform bilingual language models on Vietnamese NLP tasks, particularly text classification tasks, as demonstrated by the experiments in this study. Several other tasks such as Vietnamese Aspect Category Detection (Van Thin et al., 2021), or Vietnamese extractive multi-document summarization (To et al., 2021b) both have outperformance with monolingual BERT-based models.

Nonetheless, monolingual language models do not outperform the multilingual in all NLP tasks. More complex tasks like machine reading comprehension (Nguyen et al., 2022b) achieve better performances on multilingual pre-trained language models.

6.2 How do pre-processing techniques help improve social media text mining?

Pre-proceeding techniques are essential to improve the machine learning models significantly on social media texts, which were proven in previous studies. Nguyen and Van Nguyen (2020) proposed an approach of pre-processing data before feeding data into the model in the training stage, which was also implemented in this study to pre-process the emotion recognition task. In detail, their study is to transform emojis and emoticons, which are elements to determine the whole feeling of the content but seem to be bypassed in most normal pre-processing methods, into texts to add more detail. Their methods are appropriate pre-processing techniques based on Vietnamese social media characteristics. In our experiment, the performance of the PhoBERT model improves 5.28% from 60.16% up to 65.44% by using their approaches.

Additionally, carefully in the stage of building the dataset is also an effective way to gain optimum performances of models. The dataset ViOCD proves that the author worked well with strictly annotating and pre-processing data in building data phases. As a result, most models obtain excellent performance in identifying complaints on Vietnamese e-commerce websites.

Researchers have recently tended to apply various special approaches to optimize model performance. One of those methods is lexical normalization (van der Goot et al., 2021). As a characteristic of social media texts, unstructured content is an actual problem many models face. Most content on social media platforms is written in various formats and up to the users. Therefore, normalizing all of them into the standard in the pre-processing phase makes the model deeper understand the words and then achieve better performance in different natural language processing tasks.

6.3 How do imbalanced data impact the social media text mining?

Most datasets in social media text mining are unbalanced. Processing techniques to improve machine learning models on imbalanced datasets have recently attracted much attention from NLP researchers. As is used in this study, choosing proper metrics to evaluate models is a good method to obtain an objective view of their performances. Furthermore, if the amount of data is big enough, researchers can use the resampling method to solve the imbalance of data. The resampling technique reduces or extends the minority or majority class to achieve a balanced dataset. Besides traditional approaches to dealing with imbalanced data, new methods are developed to boost the model performance. ARCID (Abdellatif et al., 2018) and EDA (Wei and Zou, 2019) are novel methods that can address the imbalance problem in most social media datasets. This approach seeks to extract essential knowledge from unbalanced datasets by highlighting information gathered from minor classes without significantly affecting the classifier prediction performance.

7 Conclusion and Future Work

This paper described a novel evaluation benchmark for social text classification named SMTCE with four tasks emotion recognition, constructive speech detection, hate speech detection, and complaint comment detection. We implemented various approaches with BERT-based multilingual versus monolingual language models on Vietnamese benchmark datasets for each SMTC task in the SMTCE benchmark. We achieved the state-of-the-art performances: 65.44%, 84.15%, 66.43%, and 95.26% for VSMEC, ViCTSD, ViHSD, and ViOCD datasets respectively.

In future, we hope that this study will serve as a standard benchmark to develop new models on Vietnamese social media text classification or mining. Furthermore, it will motivate a range of NLP benchmarks for low-resource languages like Vietnamese.

Acknowledgement

Luan Thanh Nguyen was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2021.ThS.41.

References

Safa Abdellatif, Mohamed Ali Ben Hassine, Sadok Ben Yahia, and Amel Bouzeghoub. 2018. Arcid: A new approach to deal with imbalanced datasets classification. In *SOFSEM 2018: Theory and Practice of Computer Science*, pages 569–580, Cham. Springer International Publishing.

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 13–20, Hanoi, Vietnam. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Communications in Computer and Information Science*, pages 319–333. Springer Singapore.

Huy Duc Huynh, Hang Thi-Thuy Do, Kiet Van Nguyen, and Ngan Thuy-Luu Nguyen. 2020. A simple and efficient ensemble classifier combining multiple neural network models on social media datasets in Vietnamese. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 420–429, Hanoi, Vietnam. Association for Computational Linguistics.

Yohei Kikuta. 2019. Bert pretrained model trained on japanese wikipedia articles. `https://github.com/yoheikikuta/bert-japanese`.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 415–426. Springer.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Khang Phuoc-Quy Nguyen and Kiet Van Nguyen. 2020. Exploiting vietnamese social media characteristics for textual emotion recognition in vietnamese. In *2020 International Conference on Asian Language Processing (IALP)*, pages 276–281.

Kiet Van Nguyen, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son T. Luu, and Ngan Luu-Thuy Nguyen. 2022a. VLSP 2021 - vimrc challenge: Vietnamese machine reading comprehension. *CoRR*, abs/2203.11400.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 572–583, Cham. Springer International Publishing.

Van Kiet Nguyen, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son T Luu, and Ngan Luu-Thuy Nguyen. 2022b. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *CoRR*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021a. Monolingual vs multilingual bertology for vietnamese extractive multi-document summarization. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 555–562.

Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021b. Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 692–699, Shanghai, China. Association for Computational Linguistics.

Dang Van Thin, Lac Si Le, Vu Xuan Hoang, and Ngan Luu-Thuy Nguyen. 2021. Investigating monolingual and multilingual bertmodels for vietnamese aspect category detection.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Word Formation Processes in Masbatenyo

Romualdo A. Mabuan
Far Eastern University
Manila, Philippines
rmabuan@feu.edu.ph

Shirley N. Dita
De La Salle University
Manila, Philippines
shirley.dita@dlsu.edu.ph

Michael C. Tanangkingsing
National Taipei University of Technology
Taipei, Taiwan
miguelntut@gmail.com

Abstract

Masbatenyo (also called *Masbateño* or *Minasbate*; ISO 639-3 identifier *msb*) refers to the language spoken in the island province of Masbate in Bicol Region, the Philippines. This paper is drawn from an endogenous ethnographic study conducted by the lead author which utilized a 400,000-word corpora from written, actual spoken, and narrative data gathered from 2017-2019 in the province of Masbate. Specifically, this presentation focuses on Masbatenyo's word formation processes, which include the following: stem-based affixation (stem-forming affixes ka-, pag-, taga-/paga-, paN-, paka-, pakig-), compounding (endocentric, exocentric, synthetic, copulative), reduplication (nouns, adjectives, verbs, other lexical categories), and prefixation (uru-, Curu-). Pedagogical implications particularly in the teaching of Mother Tongue-Based Multilingual Education (MTB-MLE) are provided in light of the findings of this study.

1 Introduction

Aimed at contributing to the Philippine linguistic ecology particularly in the dearth of local literature, this study has been conducted to document the Masbatenyo language, which is a language spoken in the island province of Masbate in the Bicol Region, Philippines. Eberhard et al. (2022) classifies Masbatenyo as Austronesian, Malayo-Polynesian, Greater Central Philippines, Central Philippines, Bisayan, Central, Peripheral. It also rates Masbatenyo's language status as 3 (Wider communication) based on EGIDS or Expanded Graded Intergenerational Disruption Scale, which means that it is a language used for wider communication at various domains such as home, work, and market. Situated at the center of the Philippine archipelago and at the linguistic crossroads of surrounding islands of Bicol, Southern Tagalog, Romblon, Panay, Cebu, and Samar-Leyte, Masbate is a melting pot of dialects and cultures, making it a plurilingual/multilingual province with diverse tongues and competing grammatical systems.

2 Method

This study adopted an endogenous ethnographic study (Ginkel, 1994), which is a type of ethnography that allows a researcher to examine their own culture using an insider lens. With this approach, the primary author, as an endo-ethnographer who is a Masbatenyo native speaker, is put in a privileged position because of his a priori intimate knowledge and comprehensive view of my own culture and society. Following the tenets of ethnographic research, this approach requires methods such as field work, participant observation, interview, and examination of materials, texts, or artifacts to obtain data. Recording occurred in multiple narrative, text, voice, and video formats (DePoy & Gitlin, 2016). This paper has adapted theoretical frameworks as described in the syntactic typology of Philippine languages (Reid & Liao, 2004) and the Philippine transitivity and ergativity (Payne, 1982). Other frameworks on phonology and morphosyntactic principles helped facilitate the analysis and understanding of the morphosyntax of Masbatenyo, and these include morphosyntax (Payne, 1997), clause types Dixon and Aikhenvald's (2000), and morphosyntactic studies on Philippine languages such as nominals and noun markers (Dita, 2007), and clitics (Tanangkingsing, 2017).

3 Findings

3.1 Masbatenyo Word Formation Processes

While the authors posit that there is no clear-cut classification of word parts into nouns, adjectives, and verbs in Masbatenyo, the authors subscribe to Nolasco's (2007) observation that the interrelationships among roots, affixes, and particles and their use in discourse determine their classification. Using Nolasco's (2007) stem-based system analytical framework (also called "sapin-sapin hypothesis") he following subsections describe how affixes affect the classification of Masbatenyo root words into the different word classes such as nouns, adjectives, and verbs. Compounding and reduplication are also discussed. A root or root word is the core of a word that is irreducible into more meaningful elements (Katamba, 2006). It is the "basic element" (Fortes, 2002, p. 17), which can be left bare or to which a prefix or a suffix can attach (Kemmer, 2018). A morpheme is the smallest meaningful unit in the grammar of a language. A root contains either a free or a bound morpheme. A free morpheme can function independently as words (*kanam* 'play', *surat* 'write, letter'), while a bound morpheme appears only as part of a word, always in conjunction with a root and sometimes with other bound morphemes. Meanwhile, a stem is where the last affix is added, and it can either be in free form with only a root, or it can be in bound form with a root and a derivational stem-forming affix (Nolasco, 2005).

Bound morphemes can further be classified as derivational or inflectional morphemes. Derivational morphemes change the semantic meaing or part of speech of the affected word when combined with a root, whereas inflectional morphemes modify the tense, aspect, mood, person, or number of a verb; or the number, gender, or case of a noun, adjective, or pronoun without affecting the word's meaning or class (Baerman, 2015). Affixes are bound morphermes added to the root or stem to form a new word. Affixes can be attached before a root or stem (prefixes), after a root or stem (suffixes), within a root or stem (infixes), and before and after a root or stem (circumfixes).

Examples of derivational affixes are in the word *pagpaarado* 'plowing a filed', as used in the following example:

May	aram=ka	sa	pagpaarado?
EXI	IPERF-know=ABS.2s	OBL	IPERF-plow.a.field

'Do you know anything about plowing a field?'

The root *arado* 'plow', a noun, is a free morpheme. It becomes a bound morpheme with the addition of a causative prefix *pa–* to derive *paarado* 'to plow a field', a verb. The addition of the prefix *pag–* to this derived form *paarado* makes it a stem for the derivation of *pagpaarado* 'plowing a field', a verb and a noun. The next section discusses stem-based affixation.

3.1.1 Stem-based affixation

Stem-based analysis predicts that a word with multiple affixes will have layered structures. Nolasco (2008, 2011) proposed a simplified stem-based system to analyze voice forms in verbs with the same root or stem but with different voice affixes in Philippine languages. Nolasco posited that the stem-based analysis is a neater approach to word-formation and word analysis because it shows the formal and functional relationship between words with the same root. With stem-based analysis, attaching simultaneously the different affixes found in a word is not needed. Furthermore, the proximity or remoteness of meanings of words with the same root depends on the similarities of their stems (Nolasco, 2008).

Following Nolasco's (2011) stem-based analysis, the most productive of stem-forming affixes for verbs are *ka–*, *pag–*, *paN–*, *paka–*, and *pakig–*, which prepare the constructions for the attachment of new affixes to form new verbs. With this analysis, the root-based affixes are simplified into four voice forms: *an–*, *i–*, *–on* (the reflex of *in–* affix in Tagalog), and *-um- ~m-* (where *m–* is a replacive affix and is an allomorph of *-um-*). The *m–* replaces the first sound of *ka–*, *pag–*, *paka–*, *pakig–*, and *paN–* stems which produce the *ma–*, *mag–*, *mang–*, *maka–*, and *makig–*verbs. Employing this analysis, a verb in an intransitive construction has the voice affixes *-um- ~m,* whereas a verb in a transitive construction contains the voice affixes *–an*, *i–*, and *–on*. Aside from its application in verb formation processes, the stem-based analysis can also be applied in the analysis of noun and adjective formation.

Adopting Nolasco's (2006, 2008) stem-based analysis system, the following sections provides examples of applications of the stem-forming system of affixation in Masbatenyo to form verbs, nouns, and adjectives.

3.1.1.1 Stem-forming affixes

Stem-based affixation is evident in some Philippine languages like those presented by Nolasco (2006, 2008) such as Sorsoganon, Tagalog, Ilokano, Cebuano, and Agusan Manobo. His examples show that these languages commonly use the stem-forming affixes ka–, *pag–, paN–, paka–, and pakig–* in producing nouns, verbs, and adjectives. The following subsections provide examples of these stem-forming affixes as employed in the present data of Masbatenyo.

3.1.1.1.1 The stem-forming affix *ka–*

When the stem-forming affix *ka–* is attached to roots and other affixes, it results in the formation of nouns, adjectives, and verbs, as shown in the table below.

Table 1
The Stem-forming Affix ka–

	Affixes	Root	Gloss	Affixed	Gloss
Noun	*ka–*	ayo	'kaayo'	*ka–* ayuhan	'goodness'
		takin	'side'	*ka–* ta(ra)kin	'neighbor'
		manghod	'sibling'	*ka–* ma(ra)nghod	'siblings'
Adjective	*-um-~m*	pawa	'light, bright'	*m– +ka–* pawa	'bright'
		lisod	'difficult'	*m– +ka–* lisod	'difficult'
Verb	*-um-~m*	kadumdum	'to remember'	*n– +ka–* dumduman	'remembered'
		balyo	'to exchange'	*n– +ka–* balyuan	'exchanged'

As can be seen from Table 1, the stem-forming affx *ka–* may attach to a root to form common nouns, as shown in the following sentences:

Sin-o an imo **katakin** yana?
who DET GEN.2s neighbor now
'Who is your neighbor now?'

The *ka–* may also be attached to a suffixed root to form abstract nouns, as illustrated in (4.4). Note that the affix *ka–* is attached to *ayo* 'good' to form *kaayo* 'good', then *kaayo* is suffixed with *–an* to form *kaayuhan* 'goodness.'

Ginapangamuyo=ko gayod pirmi na sira puro la **kaayuhan**
IPERF-pray=ERG.1s PAR always LIG hope pure PAR goodness
an mag-abot sa aton buhay.
DET IPERF-come OBL GEN.1pi life
'I always pray that only pure goodness comes to our life.'

To form an adjective, the affix *ka–* may also be attached to the replacive *-um-~m*, as shown in (4.5). In this example, *kapawa* 'bright' is formed by affixing the derivational affix *ka–* to *pawa* 'brightness'. Then *m–* replaces the initial *k–* to form *mapawa* 'bright.'

Kadayaw siguro kay **mapawa** an kalangitan.
full.moon perhaps TL bright DET sky
'Perhaps it's full moon because the sky is bright.'

Following the same process, the affix *ka–* may also be attached to a root for a verb, as exemplified below with the word *nabalyuan* 'exchanged'. The word kabalyo 'exchange' comes from *ka–* + *balyo* 'exchange'. This in is attached to the suffix *an–* to form *kabalyuan* 'exchange partner' or 'something in exchange'. Finally *n–* replaces the initial *k–* to form *nabalyuan* 'exchanged.'

Baga an **nabalyuan** gad ini na akon yamit.
seem DET exchanged PAR this LIG OBL.1s clothes
'It seems like my clothes has been exchanged.'

3.1.1.1.2 The stem-forming affix *pag–*

When the stem-forming affix *pag–* is attached to roots and other affixes, it forms nouns and verbs.

Table 2
The Stem-forming Affix pag–

	Affixes	Root	Gloss	Affixed	Gloss
Noun	*-um-~m*	tuba	'coconut wine'	ma(g)nuba+R	'coconut wine maker'
Verb	*-um-~m*	pangamuyo	'prayer'	*pag–* pangamuyo, *mag–* pangamuyo	'praying'

Table 2 illustrates that the new words are formed when the replacive *m*– for nouns and the replacive *m*– for verbs are attached to derived stems. The process of forming the derived form *manunuba* ‘coconut wine maker’ starts with *pag*– + tuba ‘coconut wine’. Reduplication transforms *pagtuba* into *pagtutuba* ‘to make coconut wine’. The *m*– replaces the initial *p*– of the stem.

Manunuba an ama=ko.
coconut.wine.maker DET father=GEN.1s
‘My father is a coconut wine maker.’

Meanwhile, the stem-based affix *pag*– forms verbs that denote imperfective aspect when attached to roots, as shown in the following example.

Uupudan=mo sinda **pagpangamuyo** sa
PROS-accompany=ERG.2s ABS.3s IPERF-pray OBL
simbahan niyan?
church later
‘Will you accompany them in praying in the church later?’

3.1.1.1.3 The stem-forming affix *taga–/paga–*

When the stem-based affixes *taga*– and *paga*– are attached to roots and other affixes, they form nouns and verbs.

Table 3
The Stem-forming Affix taga–/paga–

	Affixes	Root	Gloss	Affixed	Gloss
Noun	*taga–*	Masbate	‘Masbate’	*taga–+stem* (Masbate)	‘from Masbate’
		uma	‘farm’	*taga–+stem* (farm)	‘from the farm’
Verb	*-um-~m*	siyak	‘to shout	*naga*–siyak+R	‘is shouting’
		giok	‘to thresh rice’	*naga*–giok	‘is threshing rice’

As can be seen from Table 3, *taga*– is attached to a locative noun root expressing the place of origin of a person, which could be a proper noun (e.g., *taga-Masbate* ‘from Masbate’) or a common noun (e.g., *taga-uma* ‘from the farm’).

Nag-arabot na an mga **taga-Masbate**.
IPERF-arrive PAR DET PLU from=Masbate
‘Those who are from Masbate have already arrived.’

Taga-uma ako pero dili ako maaram
from=Masbate ABS.1s CONJ NEG ABS.1s IPERF-know

maghasok kag magsanggi OBL corn
IPERF-plant CONJ IPERF-harvest sin mais.
‘I’m from the farm but I do not know how to plant and harvest corn.’

Meanwhile, when the replacive *n*– is used to replace the first sound of *paga*– stem, verbs in imperfective aspect are formed, as in *nagasiyak* ‘shouting’ from *siyak* ‘to shout’, *nagagiok* ‘threshing rice’ from *giok* ‘to thresh’, and *nagakanta* ‘singing’ from *kanta* ‘song, to sing’.

Nano kay **nagasiyak**=ka dida?
why TL IPERF-shout=ABS.2s there
‘Why are you shouting there?’

Nagagiok siya myintras **nagakanta**.
IPERF-thresh.rice ABS.3s CONJ IPERF-sing
S/He’s treshing rice while singing.

3.1.1.1.4 The stem-forming affix *paN–*

When the stem-forming affix *paN*– is attached to roots and other affixes, it forms nouns and verbs.

Table 4
The Stem-forming Affix paN–

	Affixes	Root	Gloss	Affixed	Gloss
Noun	*paN–*	tanom	‘to plant’	*paN*–(~~t~~)anom	‘plant’
		kita	‘to see, income’	*paN*–(+g)(~~k~~)ita	‘work’
Verb	*paN–*	baligya	‘to sell’	*pa*~~*N*–+(b~~)(m)aligya	‘to sell’
		bakal	‘to buy’	*pa*~~*N*–+(b~~)(+m)akal	‘to buy’

There is no change if the roots begins with the consonant *r*. Likewise, there is also no change in this affix before words starting with the consonants *d, n, s* and *t*, although the first letter of the root is dropped, as in the noun *pananom* 'plants' from *tanom* 'to plant'.

Nano an iyo mga **pananom** didi?
what DET GEN.2p PLU plant here
'What are your plants here?'

Meanwhile, the prefix *paN–* changes to *pang–* before *k* and the *k* drops, as in *kita* 'to see, income' to form the noun *pangita* 'work'.

Maayo ini na imo **pangita** kay dako an kita.
good this LIG GEN.2s work TL big DET income
'It's good that your work has a big income.'

Note that nasal assimilation occurs under the affix *paN–* where the letter *n* transforms to either *m* (i.e., from *pan–* to *pam–*) or to *ŋ* (i.e., from *pan–* to *paŋ–*) depending on the consonants following the affix, as shown in the following examples:

Maupod=ka **pagpamakal**=ko sin mga panakot
PROS-go=ABS.2s IPERF-buy=ERG.1s OBL PLU ingredient

para sa aton mga rulutuon?
CONJ OBL GEN.1pi PLU dish.to.be.cooked
'Will you go with me to buy the ingredients for our dishes?'

Ginpamaligya na ninda an inda mga kadutaan.
PERF-sell PAR ERG.3p DET GEN.3p PLU land
'They already sold their lands.'

3.1.1.1.5 The stem-forming affix *paka–*

When attached to roots and other affixes, the stem-forming affix paka– forms nouns and verbs, as shown in the table below.

Table 5
The Stem-forming Affix paka–

	Affixes	Root	Gloss	Affixed	Gloss
Noun	*-um-~m*	sala	'sin'	*m-* 'nominal affix' + (paka-+sala) + R	'sinner'
Verb	*-um-~m*	batiag	'to feel'	*m-*'verbal affix' + (paka-+batiag)	'feel'
		uli	'to go home'	*m-*'verbal affix' + (paka-+uli)	'go home'

The attachment of the *m–* 'nominal affix' to *paka–+sala* and reduplicating its first syllable results in a trait or characteristic exhibited by a person (i.e., *makasasala* 'sinner').

Makasasala an tanan na tawo sa kalibutan.
sinner DET all LIG person OBL earth
'All people on earth are sinners.'

Meanwhile, the attachment of *m–* 'verbal affix' to *paka–+batiag* and attachment to the root denotes the imperfective aspect of the verb like in *makabatiag* 'feel' from *batiag* 'to feel' and *makauli* 'go home' from *uli* 'to go home, as shown in the following sentences.

Dili=ka **makabatiag** sin sakit kun magtumar=ka
NEG=ABS.2s IPERF-feel OBL pain CONJ IPERF-take=ABS.2s
san imo bulong.
OBL GEN.2s medicine
'You can't feel pain if you take your medicine.'

Makauli na kita kay tapos na man an misa.
IPERF-go.home PAR ABS.1pi CONJ done PAR PAR DET mass
'We can now go home because the mass has already ended.'

3.1.1.1.6 The stem-forming affix *pakig–*

When attached to roots, the stem-forming affix *pakig–* forms verbs. Some examples are presented in the table below.

Table 6
The Stem-forming Affix pakig–

	Affixes	Root	Gloss	Affixed	Gloss
Verb	*pakig–*	bulig	'help'	*pakig*bulig	'to seek help'
		upod	'to get along'	*pakig*-upod	'to get along with others'
		amigo/amiga	'friend'	*pakig*-amigo/amiga	'to make friends'

Batiag=ko an iyo pakig-usad sa amon.
feel=ERG.1s DET GEN.2p IPERF-to.be.one.with.others OBL.1pe
'I can feel you're being one with us.'

3.1.1.2 Compounding

Compounding is a word formation process based on the combination of lexical elements (words or stems). This section adopts De Guzman's (2005) compound types in describing compounds in Masbatenyo. These include the following: (1) endocentric or headed compounds, (2) exocentric or headless compounds, (3) synthetic compounds, and (4) copulative compounds. The following subsections describe Masbatenyo compunds for at least three major classes of words, namely, N(oun), V(erb), and A(djective.

3.1.1.2.1 Endocentric (headed) compounds

Between the two words that usually form a compound word, one is said to function as the head and the other the non-head. Masbatenyo, just like Tagalog, is typically left-headed; that is, the category of the new compound word is typically the same as that of its headword. Depending on the category of the head, the function of the non-head in relation to its head can be determined, and given this relation, the meaning of the whole compound can typically be drawn. The following table presents some of Masbatenyo's endocentric or headed compounds.

Table 7
Masbatenyo's Endocentric (headed) Compounds

	Pattern	Example	Compound	Gloss
Noun	N + N]N	tubig 'water' + uran 'rain'	tubig-uran	'rainwater'
		bunga=n(g) 'fruit' + kahoy 'tree'	bungan kahoy	'fruit (from trees)'
	A + N]N	bag-o=n 'new' + tuig 'year'	bag-on tuig	'New Year'
	V + N]N	abri 'open' + lata 'can, tin'	abri-lata	'can opener'
Adjective	A + N]A	isip 'mind/thought' + bata 'young/child'	isip-bata	'immature'
		isip 'mind/thought' + gurang 'old/adult'	isip-gurang	'mature'
		baho=n 'smell like' + isda 'fish'	bahon isda	'fishy smelling'
		rasa=n 'smell like' + manok 'chicken'	rasan manok	'chicken tasting'
		tunga=n(g) 'half' + gab-i 'night'	tungan(g) gab-i	'midnight'
		wara=n 'nothing' + alo 'shame'	waran alo	'shameless'
		sira 'broken' + ulo 'head'	sira-ulo	'crazy'

Table 7 provides some examples of Masbatenyo endocentric compounds under noun and adjective categories. In forming an endocentric noun compound, two patterns are presented. The first pattern, N+N]N, combines two nouns (e.g., tubig 'water' + uran 'rain' = tubig-uran 'rainwater'), while the second pattern, V+N]N, combines verbs and nouns (e.g., abri 'open' + lata 'can/tin' = abri-lata 'can opener), which all produce noun compounds. Meanwhile, the combination of adjective and noun forms adjective compunds, A+N]A (e.g., isip 'mind/thought' + bata 'young/child' = isip-bata 'immature').

Masalod kita sin **tubig-uran** niyan kay
IPERF-catch ABS.1pi OBL rainwater later because
baga=n tika-uran na.
seem=DET IPERF-about.to.rain PAR

'Let's catch some rainwater later because it seems that's it's about to rain already.'

Nasira niya an **abri-lata** kaya wara lugod
IPERF-break ERG.3s DET can.opener CONJ NEG PAR
kami sin sura yana.
ABS.1pe OBL viand now
'S/He broke the can opener that's why we don't have a viand now.'

3.1.1.2.2 Exocentric (headless) compounds

Exocentric (headless) compounds have forms similar to the endocentric ones. Syntactically, most combinations appear to have a head and a modifier or a complement, but the functional relation between the constituents do not necessarily carry over to the semantic interpretation of the whole compound word. For example, a N + N form such as *balay* 'house' + *bata* 'child' does not in any way mean 'house of a child' but rather it is a nominal meaning 'uterus'. In this example, we see that the category of the supposed left head does not percolate to the compound. However, a closer analysis of the individual meanings of the two constituents together brings out the semantic content of the whole compound. In the previous example, literally the meaning is 'house of a child' which suggests the nominal meaning of 'uterus' for it is a body part that 'houses a child'. This indicates that with exocentric compounds, the semantic content transcends what the individual meanings of the constituents denote (De Guzman, 2005). The table below presents some of Masbatenyo's exocentric compounds.

Table 8
Masbatenyo's Exocentric (headless) Compounds

	Pattern	Example	Compound	Gloss
Noun	N + N]N	balay 'house' + bata 'child'	balay-bata	'uterus'
	A + N]N	patay 'dead' + gutom 'hunger'	patay-gutom	'a vagabond'
Adjective	A + V]A	bag-o=n 'new' + salta 'get.on.shore/land'	bag-on salta	'ignorant' (Lit., 'newly landed')
	N + N]A	ugali=n 'character' + hayop 'animal'	ugalin-hayop	'rude, ill-mannered'
		isip 'mind/thought' + lamok 'mosquito'	isip-lamok	'stupid'
	V + N]A	agaw 'snatch' + buhay 'life'	agaw-buhay	'dying' (Lit., 'to snatch life')

Tuna san ginbiyaan siya san iya mga
since OBL PERF-abandon ABS.3s LIG GEN.3s PLU
ginikanan baga na siya an **patay-gutom**.
Parent seem LIG ABS.3s DET vagabond
'Since s/he was abandoned by his/her parents, s/he seems like a vagabond.'

3.1.1.2.3 Synthetic compounds

Synthetic compounds are those forms in which one of the constituents is a deverbal N. One of the most common deverbalizing affixes in Masbatenyo is *pang–* which derives V into instrumental Ns, thus rendering the derived N with the meaning 'used for V-ing'. Some Masbatenyo synthetic compounds are presented in the table below.

Table 9
Masbatenyo Synthetic Compounds

	Pattern	Example	Compound	Gloss
Noun	*pang–* V]N + N]N	*pang–* patay (=pamatay) 'used.for.killing' + kagaw 'germ'	pamatay-kagaw	'germ-killer'
		pang– patay (=pamatay) 'used.for.killing' + lamok 'mosquito'	pamatay-lamok	'mosquito-killer'

Maayo ini na sabon kay **pamatay-kagaw** ini.
good this LIG soap TL germ-killer this
'This is a good soap because it's a germ-killer.'

3.1.1.2.4 Copulative compounds

Copulative compounds are those that are formed when two related words of identical categories form a compound in which not one is head and their joint meanings comprise its composite meaning. Some forms of this type are allied to exocentric type in terms of the unpredictability of their meaning. Some Masbatenyo copulative compounds are presented in the table below.

Table 10

Masbatenyo Copulative Compounds

	Pattern	Example	Compound	Gloss
Noun	N + N]N	limon 'kalamansi' + patis 'fish sauce'	limon-patis	'a mixture of kalamansi juice and fish sauce'
Adjective	V + V]A	atras 'backward' + abanti 'forward'	atras-abanti	'indecisive, ambivalent'
Verb	V + V]V	sakat 'go up' + lusad 'go down'	sakat-lusad	'to move up and down'
		unlod 'sink' + lutaw 'appear'	unlod-lutaw	'to appear and disappear'
		guwa 'go outside' + sulod 'go inside'	guwa-sulod	'to go outside and inside'

Kamakabaradli gani sani na mga bata sin
annoying PAR this LIG PLU child ERG
kakaguwa-sulod.
IPERF-go.outside.and.inside

These kids are really annoying for going outside and inside repeatedly.'

3.1.1.2.5 Reduplication

Reduplication is a very distinct characteristic of Austronesian languages including the Philippine-type languages. Hurch (2005) defined reduplication as a morphological process in which the root or stem of a word or part of it or even the whole word is repeatedly exactly or with a slight change. Nevins and Vaux (2003) noted that reduplication is found in a wide range languages and language groups around the world, though its level of linguistic productivity varies. It is often used when a speaker adopts a tone more "expressive" or figurative than ordinary speech and is also often, but not exclusively, iconic in meaning. In Masbatenyo, reduplication is probably one of the most dominant and interesting features. It occurs among nouns, verbs, adjectives, adverbs, and also numerals. Furthermore, the wide range of reduplication signals various meanings such as plurality, intensity, iterativity, frequency, limitation, inter alia.

Some examples of reduplication among word classes in Masbatenyo are presented in separate tables below. Note that the discussion of these word classes are subsumed in their own chapters in this dissertation.

3.1.1.2.5.1 Reduplication in Nouns

Masbatenyo noun reduplication usually indicates plurality. The reduplicant shape depends on the CV morphological structure of the base word. Some examples of the reduplicated nouns are presented below.

Table 11

Reduplication in Nouns

Base	Gloss	Reduplicated	Gloss
istakan	'container'	*uru*-istakan	'mini/toy container'
ingkudan	'chair'	*uru*-ingkudan	'mini/toy chair'
baruto	'boat'	*buru*-baruto	'mini/toy boat'
platito	'saucer'	*puru*-platito	'mini/toy saucer'
balay	'house'	*balay*-balay	'small house'; 'households'
ido	'dog'	*ido*-ido	'toy dog'

3.1.1.2.5.2 Reduplication in Adjectives

Reduplication in adjectives encodes intensity, moderation, and superlativity. Consider the following examples.

Table 12
Reduplication in Adjectives

Base	Gloss	Reduplicated	Gloss
maputi	‘white’	maputi-puti	‘very white’
madulom	‘dark’	madulom-dulom	‘very dark’
manamit	‘delicious’	manamit-namit	‘very delicious’
maniwang	‘thin’	maniwang-niwang	‘very thin’
maarat	‘salty’	maarat-*arat*	‘somewhat salty’
maaslom	‘sour’	maaslom-*aslom*	‘somewhat’
matulin	‘fast’	matulin-*tulin*	‘very fast’
mahinay	‘slow’	mahinay-*hinay*	‘very slow’

3.1.1.2.5.3 Reduplication in Verbs

Reduplication in verbs signals the number of agents, iterativity of the action, and frequency of action. The table below presents some examples of these verbs.

Table 13
Reduplication in Verbs

Base	Gloss	Reduplicated	Gloss
lakat	‘walk’	nag*lakat*-lakat	‘keeps on walking’
inom	‘drink’	nag-*inom*-inom	‘keeps on drinking’
lumpat	‘jump’	Naglumpat-lumpat	‘kept on jumping’
tawa	‘laugh’	nag*tawa*-tawa	‘keeps on laughing’
balik	‘move back’	nag*b*in*alik*-balik	‘kept on coming back’
sayaw	‘dance’	nag*sayaw*-sayaw	‘keeps on dancing’
surat	‘write’	nag*s*in*urat*-surat	‘kept on scribbling’
hikap	‘touch’	nag*hikap*-hikap	‘keeps on touching’

3.1.1.2.5.4 Reduplication in Other Lexical Categories

Reduplication in adverbs may also occur in Masbatenyo. The process seems to encode intensity and emphasis. Consider the following examples.

Table 14
Reduplication in Adverbs

Base	Gloss	Reduplicated	Gloss
adlaw	‘day’	*adlaw*-adlaw	‘every day’
bulan	‘month’	*bulan*-bulan	every month’
dali	‘immediate’, ‘hurry’	*dali*-dali	‘immediately’, ‘hurriedly’, ‘haphazardly’
ungod	‘diligent’	*ungud*-ungod	‘diligently’
sutoy	‘continuous’	*sutoy*-sutoy	‘non-stop’

Numerals also exhibit reduplication, which primarily encodes ordinal or sequential and limitative meaning. The examples below illustrate this.

Table 15
Reduplication in Numerals

Base	Gloss	Reduplicated	Gloss
isad/usad	‘one’	*isad*-isad/ *usad*-usad	‘one by one’, ‘individually’
duha/duwa	‘two’	*duha*-duha/ *duwa*-duwa	‘by pairs’

tulo	'three'	*tulu*-tulo	'by threes'
upat	'four'	(tag-)*upat*-upat	'four each'
pito	'seven'	(tag-)*pitu*-pito	'seven each'

4 Conclusion

This paper discussed some important features of Masbatenyo morphology. These include the word formation processes including affixation, compounding and reduplication process that happens in some Masbatenyo lexical categories such as nouns, adjectives, verbs, and other lexical categories such as adverbs and numerals. The findings offered may be used to provide richer characterization of the language, and as a reference in the preparation of instructional materials for the mother tongue-based multilingual education (MTBMLE), and for the creation of Masbatenyo orthography. As the discussion presented here is not exhaustive, future research may further explore the complex affix system of Masbatenyo across its dialects, particularly the semantic roles that are cross-referenced by the affixes.

References

Baerman, M. (2015). *The morpheme*. Oxford University Press.

DePoy, E., & Gitlin, L. N. (2016). *Introduction to research: Understanding and applying multiple strategies.* ScienceDirect.

Dita, S. N. (2007). *A reference grammar of Ibanag* (Doctoral dissertation). De La Salle University, Manila, Philippines.

Dixon, R. M. W., & Aikhenvald, A. Y. (2000). Introduction. In R. M. W. Dixon and A. Y. Aikhenvald (Eds.), *Changing valency: Case studies intransitivity* (pp. 1-29). Cambridge University Press.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.) (2022). *Ethnologue: Languages of the world*. Ethnologue.

Fortes, F. C. L. (2002). *A constraint-based morphological analyser for concatenative and non-concatenative morphology of Tagalog verbs*. MS Thesis. De La Salle University, Manila, Philippines.

Ginkel, R. V. (1994). Writing culture from within: Reflections on endogenous ethnography. *ETNOFOOR, 7*(1), 5-23.

Katamba, F. (2006). *Morphology* (2nd ed.). Palgrave Macmillan.

Kemmer, S. (2018). *Words in English: Structure.* http://www.ruf.rice.edu/~kemmer/Words04/structure/index.html

Nolasco, R. M. (2006). *Ano ang S, A, at O sa mga Wika ng Pilipinas?* Paper read at the 9th Philippine Linguistics Congress, January 2006. University of the Philippines, Diliman, Quezon City.

Nolasco, R. M. (2008, July). *The prospects of multilingual education and literacy in the Philippines*. Paper presented at the 2nd International Conference on Language Development, Language Revitalisation and Multilingual Education. Bangkok, Thailand.

Nolasco, R. M. (2011). *Grammar notes on the national language Filipino* (Unpublished draft). University of the Philippines, Quezon City.

Payne, T. E. (1982). Role and reference related subject properties and ergativity in Yup'ik Eskimo and Tagalog. *Studies in Language, 6*(1), 75-106.

Payne, T. E. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge University Press.

Reid, L., & Liao, H. (2004). A brief syntactic typology of Philippine languages. Language abd Linguistics, 5(2), 433-490. https://scholarspace.manoa.hawaii.edu/bitstream/10125/32991/1/A55.2004.pdf

Tanangkingsing, M. (2009). *A functional grammar of Cebuano* (Doctoral dissertation). National Taiwan University, Taipei, Taiwan.

Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and Combinatorial Feature Embedding Based on CNN, LSTM and BERT

Manel Affi
RIADI / ENSI
University of Manouba / Tunis
Tunisia
manel.affi@ensi-uma.tn

Chiraz Latiri
LIPAH / FST
University of Tunis El Manar/ Tunis
Tunisia
chiraz.latiri@gnet.tn

Abstract

One of the most important factors that significantly affects the quality of sequence labeling models is the selection and encoding of input features. The complexity and morphological richness of Arabic language is the main reason why most existing Arabic Named Entity Recognition (NER) systems rely heavily on hand-crafted feature engineering. To overcome such a limitation, we propose a novel neural network architecture to tackle the Arabic NER task. The proposed approach takes advantage of the recent success of deep neural networks in various Natural Language Processing (NLP) applications. We use a variant deep neural network architecture combined with a combinatorial feature embedding based on Convolutional Neural Network (CNN), Long Short-Term Memory networks(LSTM) and BERT to generate rich semantic and syntactic word representation vectors. Without using any external knowledge or hand-crafted feature engineering, the proposed models outperform the state-of-the-art systems on the ANERCorp dataset by yielding an F1-score of 93.34% and 93.68 % using bidirectional LSTM-CRF (BLC) and bidirectional GRU-CRF (BGC) architectures, respectively.

1 Introduction

Named Entity Recognition (NER) is defined as an information extraction task which aims to identify, extract and automatically categorize named entities into a set of predefined categories such as persons, organizations, locations, etc (Nadeau and Sekine, 2007). It is also considered as a mandatory preprocessing module in many wider NLP applications, such as information retrieval, speech recognition, syntactic parsing, machine translation, text classification, question answering, entity coreference resolution and entity linking and disambiguation.

There is a fair amount of literature on NER research in English, Chinese and other widely spoken languages. The massive growth of Arabic content on the web has led to increased demand for developing accurate and robust Arabic NLP tools. In recent years, Arabic NER has become a challenging task and has received increasing attention from current researchers due to its characteristics and peculiarities (Farghaly and Shaalan, 2009), and given the limited availability of annotated datasets (Zirikly and Diab, 2015).

Researchers interested in Arabic NER have mainly pursued three approaches: rule-based (Mesfar, 2007; Zaghouani, 2012), machine learning-based (Benajiba and Rosso, 2008) and hybrid methods (Abdallah et al., 2012; Oudah and Shaalan, 2017). All these three methods suffer from the same problem, as it requires a lot of language-specific knowledge and feature engineering to obtain useful results. This is even more highlighted by the deficiency of linguistic resources and the complex morphology and syntax of the language.

Recently, the deep learning paradigm (Mishra and Gupta, 2017) has emerged and led to impressive advances in fields such as speech processing (Oord et al., 2016) and image recognition (Hu et al., 2018). For NLP, the application of deep learning has proven to be very effective as it yields the state-of-the-art in various common NLP tasks such as sequence labeling (Ma and Hovy, 2016) and named entity recognition (Affi and Latiri, 2021) for the English language. Unlike traditional methods, deep learning is an end-to-end model that does not rely on data pre-processing, manual feature extraction or large amounts of task-specific resources, and it can be applicable to different languages and domains. This makes it an attractive solution for complex and low resource languages like Arabic.

Motivated by the success of deep learning in many NLP tasks, we propose a novel Arabic NER approach based on deep neural networks. The proposed model takes benefits from CNN and LSTM to induce character-level representations of words. These representations are fed in conjunction with BERT word embeddings to a deep learning model for further sequence modeling. Two RNN models are investigated in this work: the Bi-LSTM and Bi-GRU models. Finally, we use a Conditional random fields (CRF) layer to get the probability distribution over the tags.

The main contributions of the proposed Arabic NER model are summarized as follows:

- We propose a novel neural network architecture based on variant deep neural network architectures combined with word embeddings based on the CNN, LSTM and BERT to tackle the Arabic NER task.
- We study the impact of the combinatorial feature embedding based on the CNN, Bi-LSTM and BERT on Arabic NER.
- We investigate the use of two types of character-level representations to slove the Arabic NER task. We also show that using pre-trained word embeddings enhances the system performance.
- We give an empirical evaluation of this system on the ANERCorp dataset.
- We evaluate our model against different baselines to demonstrate the empirical strength of our work.

The remainder of this paper is structured as follows. In the next section, we review the related work briefly. Section 3 presents the proposed model architecture in detail. Section 4 describes the training mechanism. Section 5 presents the experimental results and discussions. Finally, we conclude and discuss possible improvements for future work in section 6.

2 Related work

Recent NER research studies have used deep learning, which proves its powerful ability for feature abstraction. As far as Arabic NER is concerned, some consideration has been specified to neural models. (Mohammed and Omar, 2012) proposed one of the first artificial neural network for Arabic NER. The authors introduced a simple feed forward ANN to tackle the task. The suggested model consists of three steps: data pre-processing, transforming Arabic letters to Roman letters, after that the collected data is categorised using a neural network. This approach was tested on the ANERCorp dataset. The obtained results showed that the proposed method gave better performance than the decision trees, and the model accuracy improved with the size of the data. A tagging model for Arabic NER was proposed by (Awad et al., 2018). To solve the out-of-vocabulary (OOV) problem, a CNN character-based embedding layer was concatenated with Word2Vec word embeddings to initialize the word representations. Then, the vector representations were fed to a Bi-LSTM-CRF architecture. The system was evaluated on the ANERCorp and AQMAR datasets. Many hyper-parameter setting were checked, and the model achieved an F1-score of 75.68% . Recently, the attention mechanism has demonstrated its effectiveness in different NLP applications such as machine translation. It can handle long input sequences and learn to focus only on the most important words. In (Ali et al., 2018), the authors added a self-attention layer to enhance their Arabic NER model. The proposed model learnt to give low or high attention to words based on their context in the input sentence. Also, the writers in (Ali et al., 2019) introduced a multi-attention based model for Arabic NER. This proposed system showed the impact of using word-level embeddings and character-level representations followed by a Bi-LSTM and self-attention layers. Tested on the ANERCorp dataset, it achieved an F1-score of 91.31$\%$.

A recent work (Al-Smadi et al., 2020) examined the impact of transfer learning on a Pooled-GRU architecure for Arabic NER. Their model outperformed the Bi-LSTM-CRF model suggested by (Sa'a et al., 2018).

3 Proposed neural network architecture

In this section, we describe the architecture of our neural network model based on BERT word embeddings and CNN and LSTM-based character level vectors for the Arabic NER task. The proposed system comprises three main layers, namely an word embedding layer, a context layer, and a tag decoder layer.

Definition: Assuming an input sentence S coming from a series of tokens of size $\parallel V \parallel$ with a sequence of labels $Y = (y_1, ..., y_n)$, its word-level input is defined as $X = (x_1, ..., x_n)$, where x_i is the i^{th} token in a series of words of size $\parallel n \parallel$, and n is the number of the tokens in the input sentence. With a neural sequence labeling system, the objective is to find the adequate entity labels for all tokens in X, and then assign a sequence of annotations $\tilde{y} \in Y$ to it, where Y is the list of all possible tag categories. The highest probable symbol sequence $\tilde{y}$ is outputted by maximizing the sequence posterior probability of an optimal resulting sequence $\tilde{y} = (\tilde{y_1}, \tilde{y_2}, ..., \tilde{y_n})$, which closely matches the gold annotations sequence $y = (y_1, y_2, ..., y_n)$ that indicates the right label for each given token.

As shown in Figure 1, our proposed model is composed of three main blocks which are explained below.

3.1 Embedding layer

The representation layer is designed to provide the main features that will be used as a key component for our NER system. The quality of features has a significant impact on the model performance. Traditionally, features are hand-crafted keeping some interesting rules that may not be relevant to other areas. Thus, many state-of-the-art NLP approaches tend to use various deep neural networks architecture for outputting distributed word representations in order to catch both syntactic and semantic patterns of words. In distributed embeddings, the model is generalizable because each word refers to dense low-dimensional real-valued vectors in the space, so that tokens with approximated semantic and syntactic properties will have similar vector representations. However, learning high-quality word vectors can be challenging. Ideally, they should represent successfully both complex characteristics of word uses and how these uses change according to their linguistic context. Hence, word embeddings have attracted a lot of consideration from many researchers (Mikolov et al., 2013; Lai et al., 2016).

Over the last years, different tools, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been widely used in the NLP field. In order to reach high-quality of word embeddings, researchers have recently introduced different techniques to generate various embeddings for the same word depending on its context (Devlin et al., 2018;

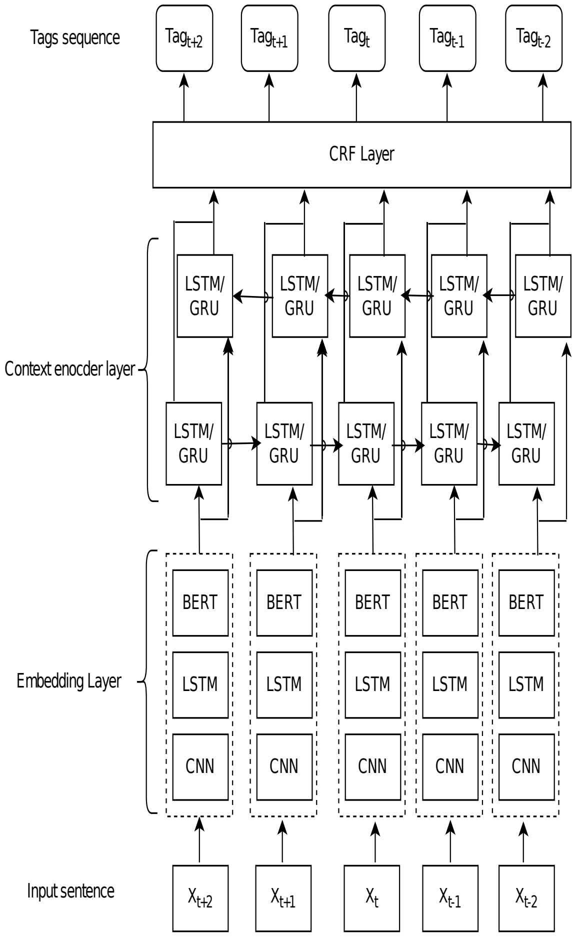


Figure 1: Main model architecture

Peters et al., 2018).

However, using a word embedding alone as the smallest feature representation unit may result in some loss of important information. For morphologically rich languages, such as Arabic, we need to catch all morphological and orthographic patterns. As the word embedding encodes semantic and syntactic word relationships, character-level representations model important morphological and shape information. Inspired by this integration, we acquire the word representations from BERT word embeddings and two different character-level representations extracted from a CNN and LSTM.

3.1.1 Word embedding layer

In distributed embeddings, words refer to a vector in a continuous space to capture syntactic and semantic relations among them. In this work, we use BERT (Devlin et al., 2018) as distributed word embeddings. BERT was published by researchers at Google AI Language in late 2018. Its full name is Bidirectional Encoder Representations from Transformers. It represents a language model representation based on self-attention blocks. The main innovation of this model is the pre-training approach, which determines word and sentence-level representations based on masked-language modeling and next sentence prediction training. BERT is pre-trained in various languages using existing unlabeled data. The pre-trained deep bidirectional model with one output layer has become a state-of-the-art in many NLP applications such as multi-genre natural language inference, named entity recognition and question answering. The idea is to have a common architecture adequate for many NLP applications and a pre-trained model that decreases the need for labeled data and boost the performance for different downstream NLP tasks.

To obtain the pre-trained contextual word embedding for Arabic from BERT model, we make use of a publicly available pre-trained resource for research purposes: AraBERT[1] (Antoun et al., 2020), with a dimension of 768 and trained on 70M sentences with 3B words of Arabic text. AraBERT uses the same Google's BERT architechture and BERT-Base[2] configuration with 768 hidden layers, 12 layers of transformer encoders and 12 attention heads. For each token, a contextual word embedding is generated by averaging the corresponding word, subword and position embeddings in the last four layers.

3.1.2 Character embedding layer

In addition to word embeddings, character-level embeddings are also used to represent input words. It is used to handle OOV words that are not present in the trained word vectors. Especially for the Arabic language which is rich in morphology and syntax, character-level embeddings help to efficiently extract morphological patterns of each token. We use two different neural networks, the CNN and LSTM, to extract the character-level patterns.

Character-level CNN model

Among deep learning techniques, the CNN is a well-known architecture that is widely used to capture local information. CNNs, which were originally used for image processing (Krizhevsky et al., 2012), are now also used to capture character-level information for various NLP tasks. The authors in (Labeau et al., 2015) and (Santos and Guimaraes, 2015) successfully studied the use of CNNs to model character-level features in POS-tagging and NER tasks, respectively. The writer in (Collobert et al., 2011) also used CNNs to improve the semantic role labeling task. The authors in (Kim et al., 2016) proposed an interesting CharCNN model, which is a character-aware neural language model that learns character-based word representations using CNNs.

In this approach, we also investigate the use of the CNN architecture to represent each character in the input token by a character vector. First, each token is mapped to a suitable representation using its character embeddings. This character embedding vectors are then passed through a convolutional layer using multiple filters to detect various features. Formally, for an input token W with T characters $\{ w_1, w_2, ..., w_T\}$, matrix $\Omega \in R^{d_{ch}*T}$ is learned to map the token into character embeddings, where d_{ch} is the dimension of the character representations.

[1] `https://huggingface.co/aubmindlab/bert-base-arabertv01`

[2] More details about the transformer architecture can be found in (Vaswani et al., 2017)

The convolution layer has as input the sequence of character embeddings $\{ \omega_1, \omega_2, ..., \omega_T\}$. Then, a convolution operation is used between Ω and filter $H \in R^{d_{ch} * w}$ of width h. After that, we add a bias and apply the $tanh$ function to add non-linearity to acquire the feature map $f \in R^T$:

$$f_i \quad = \quad tanh(< \Omega_{i:i+h-1}, H > +b) \tag{1}$$

where $< \Omega_{i:i+h-1}, H >$ denotes the Frobenius inner product and b represents the bias term.
Finally , a max-pooling operation (Eq 2) is applied to learn a single feature for all feature maps.

$$Q_H^W \quad = \quad max_i f_i \tag{2}$$

where Q_H^w represents the character-level embedding of a token W produced by filter H to capture the local information.

Character-level LSTM model

LSTM models are also used to extract character-level patterns. LSTM is a well-recognized scheme of learning long-term dependencies. The typical LSTM unit employs a memory cell controlled by a forget gate that determines which previous memory should be scaled into the next time-step. Similarly, the new input to memory cells is passed through an input gate.
Formally, the implementation of the LSTM memory unit is described as follows:

$$f_t \quad = \quad \sigma(W_f . [x_t, h_{t-1}] + b_f) \tag{3}$$

$$i_t \quad = \quad \sigma(W_i . [x_t, h_{t-1}] + b_i) \tag{4}$$

$$\tilde{c}_t \quad = \quad tanh(W_c . [x_t, h_{t-1}] + b_c) \tag{5}$$

$$c_t \quad = \quad f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

$$o_t \quad = \quad \sigma(W_o . [x_t, h_{t-1}] + b_o) \tag{7}$$

$$h_t \quad = \quad o_t \odot tanh(C_t) \tag{8}$$

where σ is the logistic sigmoid function, $\odot$ represents the element-wise multiplication, f, i, o and c are respectively the forget gate, the input gate, the output gate and the cell vector, x_t represents the input vector at the t^{th} time-step, h_t is the hidden layer state vector that saves all the important information at the t^{th} time-increment, W_f, W_i, W_c and W_o denote the weight matrices of the various gates, and b_f, b_i, b_c and b_o represent the bias vector.

3.2 Context encoder layer: Bi-LSTM/Bi-GRU

The context layer aims to get the local dependencies using neighboring words for every word. The local context is crucial for effectively predicting labels, since there could exist strong relationships among neighboring tokens in a sentence. Therefore, it is important to model the local context information for each word. In the NER task, it is common to use the RNN as a context encoder model. It treats the input sequence in order, where shuffling or reversing the time-steps influences the extracted representations from the input sequence. The longer the input sequence, the less precise the RNN becomes because it is difficult for the network to remember the output of the time steps far away from the previous (Goyal et al., 2018). This issue is named the vanishing gradient problem. LSTM and GRU are variations of the RNN that help in dealing with the vanishing gradient issue and can learn long dependency input. In the NER task, both previous and future information is helpful for prediction. For that, we use the Bi-LSTM and Bi-GRU networks as a context encoder layer after obtaining word embeddings from the concatenation (Eq.9) BERT model, CNN as well as the LSTM-based character embeddings B^{emb}, CN^{emb} and CL^{emb}, respectively.

$$X_t^{emb} = B_t^{emb} \oplus CN_t^{emb} \oplus CL_t^{emb} \tag{9}$$

where B_t^{emb} is is the embedding learned by BERT, CN_t^{emb} is the CNN-based character embedding, CL_t^{emb} represents the LSTM-based character embedding, and X_t^{emb} is the vector representation for the t^{th} word of the input sentence.

3.2.1 Bi-LSTM

In addition to learning character embedding, we use LSTM, which is formally explained in section 3.1.2 also to learn the contextual features of a sequence of words. To better represent the current word information, it is beneficial to obtain both prior and posterior contexts. In our model, we use Bi-LSTM to effectively predict the current label information. As input, Bi-LSTM takes the vector representation of each word X_t^{emb}. The basic idea is to present the sequence from left to right using a forward layer (Eq.10) and to compute a representation of the same sequence in reverse via a backward layer

(Eq.11). These two different networks use various parameters to learn privious and future patterns. The left and right context representations are then concatenated to form the final output (Eq.12)

$$\overrightarrow{h_t} = f(\overrightarrow{h_{t-1}}, [X_t^{emb}]) \quad (10)$$

$$\overleftarrow{h_t} = f(\overleftarrow{h_{t+1}}, [X_t^{emb}]) \quad (11)$$

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (12)$$

where $f(.)$ represents the unidirectional LSTM cell, $\overleftarrow{h_t}$ and $\overrightarrow{h_t}$ represent respectively the hidden states of the backward LSTM cell and the forward cell, and t is the index of the encoding steps.

3.2.2 Bi-GRU

The GRU was proposed by (Cho et al., 2014). Similarly to the LSTM cells, the GRU was modeled to adaptively update or reset its memory content by using reset and update gates which are similar to the forget and input gates of the LSTM unit. The update gate decides what information to use as input at the next time step, whereas the reset gate uses the previous time step output to decide which information should be removed and which information is useful for the current time step input. The update gate, the reset gate, and the hidden state h_t at time step t are updated utilizing the equations described below.

$$u_t = \sigma(Z_u.[x_t, h_{t-1}] + b_u) \quad (13)$$

$$r_t = \sigma(Z_r.[x_t, h_{t-1}] + b_r) \quad (14)$$

$$\tilde{h}_t = tanh(Z.[x_t, h_{t-1}.r_t] + b) \quad (15)$$

$$h_t = (1 - u_t).h_{t-1} + u_t.\tilde{h}_t \quad (16)$$

where σ denotes the logistic sigmoid function, u, r and $\tilde{h}$ respectively represent the update gate, the reset gate and the cell state, x_t is the input vector at time t, h_t represents the hidden layer state vector at time t, Z_u, Z_r and Z denote the weight matrices of the different gates and the cell state, respectively, and b_u, b_r and b represent the bias vector. In this work, we use a Bi-GRU to process the input sentence by a forward layer from left to right (Eq.17) and from the other side by a backward layer (Eq.18). Then, the forward layer hidden state and the backward layer hidden state are combined to represent the final hidden state (Eq.19).

$$\overrightarrow{h_t} = g(\overrightarrow{h_{t-1}}, [X_t^{emb}]) \quad (17)$$

$$\overleftarrow{h_t} = g(\overleftarrow{h_{t+1}}, [X_t^{emb}]) \quad (18)$$

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (19)$$

where $g(.)$ represents unidirectional GRU unit, $\overleftarrow{h_t}$ and $\overrightarrow{h_t}$ respectively denote the hidden states of the backward GRU unit and forward unit, and t is the index of the encoding steps.

3.3 Tag encoder layer: CRF layer

For sequence tagging tasks, it is beneficial to take into consideration the relationship between tags in neighborhoods; and for a given input sentence, it is important to jointly decode the label chain that yields the best resulting label sequence. In NER task with the IOB annotation, there is a strong dependency between labels; e.g., I-PER cannot follow I-ORG. Therefore, modeling the label correlations is important for the NER task (Ma and Hovy, 2016; Liu et al., 2018). Following (Ma and Hovy, 2016) and (Huang et al., 2015), we incorporate a CRF (Lafferty et al., 2001) layer upon the context layer to jointly capture the label correlations.

Formally, for a given input word sequence $X = (x_1, x_2, ...x_n)$ with a sequence of annotations $Y = (y_1, y_2, ...y_n)$, we denote Ω as the scoring matrix resulting from the context layer. The size of matrix Ω is n*m, where n and m denote the length of the sentence and the number of different labels, respectively. $\Omega_{i,j}$ is the score of the j^{th} label of the i^{th} word in the input sentence. We also use $Y_{(X)}$ to define the set of possible label sequences for X.

The score of a given sentence is determined through Eq.20:

$$S(X, y) = \textstyle\sum_{i=0}^{n} \psi_{y_i, y_{i+1}} + \sum_{i=1}^{n} \Omega_{i, y_i} \quad (20)$$

where ψ denotes the matrix of transition scores. The transition score matrix values are learned during training. After that, a softmax function is applied to regularize the conditional probability of the output path y:

$$P(y|X) = \frac{\exp^{S(X,y)}}{\sum_{\tilde{y} \in Y_{(X)}} \exp^{S(X,\tilde{y})}} \quad (21)$$

During the training step, the aim is to maximize the following log-probability of the right label sequence:

$$\log(P(y|X)) = S(X, y) - \log(\textstyle\sum_{\tilde{y}} \exp^{S\ (X,\tilde{y})}) \quad (22)$$

For the evaluation step, we aim to retrieve the most probable output label sequence y^* by maximizing the score function:

$$y^* = \arg\max_{\tilde{y} \in Y_{(X)}} S(X, \tilde{y}) \quad (23)$$

Finally, we employ the Viterbi (Forney, 1973) algorithm to train the CRF layer and decode the best sequence y^*.

4 Training mechanism

4.1 Dataset description

To prove the effectiveness of our Arabic NER system, we conduct extensive experiments using the ANERCorp[3] which was developed by Benajiba (Benajiba et al., 2007) from diverse online resources and which is freely available for research purposes. It has 4,901 sentences with 150,286 tokens of articles from Modern Standard Arabic. The ANERCorp dataset is manually annotated where each word was tagged with one of the following tags: person, location, company, and others. ANERCorp is corpus that has two corpora, namely training and testing corpora.

4.2 Evaluation metrics

In order to evaluate the efficiency of our models, the standard measures for NER precision (P), recall (R), and F1-Score (F) are computed.

The precision measure represents the percentage of name entities found by the system and which are correct. It can be computed as:

$$P = \frac{NumberOfCorrectNamedEntities}{NumberOfNamedEntitiesExtracted} \quad (24)$$

On the other hand, the recall measure represents the percentage of name entities that the system extracted correctly out of the overall number of named entities existing in the corpus, and it can be expressed as:

$$R = \frac{NumberOfCorrectNamedEntities}{TotalNumberOfNamedEntities} \quad (25)$$

Finally, the F1-score is the most commonly used and is computed based on the precision and recall. This is defined as:

$$F1 - score = 2X\frac{P * R}{P + R} \quad (26)$$

4.3 Hyper-parameter settings

Our neural network model is implemented using Keras environment and TensorFLOW API. A word vector embedding represents the input of the system. We obtain the word vector using the concatenation of the pre-trained BERT word vector of dimension 768, CNN and LSTM-based character level representations where each of them is of dimension 20. The training is performed using the backpropagation algorithm to update the parameters of all models during the training process. We choose the Adam algorithm (Kingma and Ba, 2014) and we set a learning rate of value 1e-5. Our model uses two layers of the Bi-LSTM and Bi-GRU networks with hidden layer nodes of dimension 512. For the over-fitting problem, a 20% dropout is regarded as an additional measure to control the input of the neural model and alleviate the over-fitting issue very well. In each iteration, we split the entire training data into batches and pass one batch at a time. In our experiments, we train our model with a batch size of 64 for only three epochs.

5 Experimental results and discussions

In this section, we present the experimental results obtained by using different architectures based on two RNN variants, namely the Bi-LSTM and the Bi-GRU applied on ANERCorp dataset. We run three parts of experiments: The first part focuses on the selection of the best architecture giving the highest performance. In the second part, we discuss the obtained results, while the third part of the experiments presents a comparison between the best-selected architecture and the best previous state-of-the-art methods.

5.1 Results

To explore the performance of our proposed models and show the effect of different architectures. We have conducted a series of comparative experiments. We have examined five different word embedding choices: Randomly initialized Word Embeddings(RWE), BERT word embeddings, two different types of character embeddings based on the

[3] https://camel.abudhabi.nyu.edu/anercorp/

CNN and LSTM and CNN-LSTM-BERT (CLB) combinatorial feature embeddings. All these word embeddings are tested in combination first with BLC and then with BGC models. Tables 1 and 2 show the obtained results.'Exp', 'P', 'R', and 'F1' denote experiment, precision, recall and F1-score, respectively.

Exp	Model	R	P	F1
1	RWE-BLC	85.40	84.91	85.15
2	CNN-BLC	90.62	90.43	90.52
3	LSTM-BLC	90.50	90.30	90.39
4	BERT-BLC	92.33	92.90	92.61
5	CLB-BLC	**93.20**	**93.50**	**93.34**

Table 1: Results obtained based on Bi-LSTM.

Exp	Model	R	P	F1
1	RWE-BGC	85.69	85.09	85.38
2	CNN-BGC	90.97	90.75	90.85
3	LSTM-BGC	90.80	90.57	90.68
4	BERT-BGC	92.58	93.10	92.83
5	CLB-BGC	**93.60**	**93.77**	**93.68**

Table 2: Results obtained based on Bi-GRU.

Related to the experimental results presented in Tables 1 and 2, we can observe that the combination of BGC and CLB combinatorial word embeddings exhibits the highest performance and outperforms the model based on BLC which reaches 93.20% precision, 93.50% recall and a 93.34 F1-score% by achieving 93.77% precision, 93.60% recall and a 93.68% F1-score. In table 1 and 2, In five different experiments, the obtained results show that the Bi-GRU performs better than Bi-LSTM by about 0.34% in the F1-score with the CLB combinatorial feature embeddings.

5.2 Effects of different word embeddings

To analyze the effects of different types of word embeddings, we conduct experiments using the BLC and the BGC with five different combinations of embeddings. The results are presented in Tables 1 and 2. In experiment 1, the model applies only the RWE. In experiments 2 and 3, the models use embeddings that combine one type of character-level embedding, the CNN and LSTM, respectively. In experiment 4, the models use pre-trained BERT word embeddings whereas in experiments 5, the models use a combinatorial feature based on the CNN, LSTM and BERT. Experiment 1 shows that the RWEs exhibit the lowest performance. The results of experiments 2 and 3 significantly outperform those of experiment 1, indicating that character-level embedding is useful for handling OOV words in Arabic NER tasks. Additionally, in experiments 2 and 3, the CNN model outperforms the LSTM model in extracting character-level features by 0.12% and 0.17% in the F1-score using BLC and BGC models, respectively. In experiment 4, we investigat the effect of using pre-trained BERT word embeddings. It demonstrates that pre-trained BERT embeddings have the most positive impact on the performance by achieving 92.33% and 92.58% F1-score using the BLC and BGC, respectively. In experiment 5, the proposed models utilizing all three types of embedding (CNN,LSTM and BERT) for word representation exhibit the highest performance achieving an F1-score of 93.34% 93.68% F1-score using BLC and BGC architectures, respectively. This indicates that in Arabic NER, the use of character-level based word representations and Arabic pre-trained word embeddings are effective for handling OOV words and effciently capturing semantic and syntactic word relationships.

5.3 Performance comparison with state-of-the-art methods

In this part of these experiments, we compare the performance of our proposed models with the models used in previous studies on ANERCorp. Table 3 presents the results of three competitive Arabic NER models from the studies of (Alsaaran and Alrabiah, 2021)(BERT-GRU), (Ali et al., 2019) (Multi-Attention) and (El Bazi and Laachfoubi, 2019) (CW-Bi-LSTM-CRF) evaluated on entity-level matching.

The authors in (Alsaaran and Alrabiah, 2021) introduced a novel Arabic NER model for Arabic based on the BERT and the Bi-GRU. This model achieves a 92.28 % F1-score on ANERCorp. In addition to word-level embeddings, the writers in (Ali et al., 2019) adopted character-level embeddings and combined them via an embedding-level attention mechanism followed by Bi-LSTM and a self-attention layer. This model yielded an F1-

Model	R	P	F1
BERT-BGRU	92.40	92.20	92.28
Multi-Attention	90.62	90.43	90.52
CW-Bi-LSTM-CRF	-	-	90,60
CLB-BLC (**our**)	**93.20**	**93.50**	**93.34**
CLB-BGC (**our**)	**93.60**	**93.77**	**93.68**

Table 3: Comparison between our best models and three Arabic NER systems on ANERCorp dataset.

score of 91.31% on ANERCorp. A Bi-LSTM-CRF based neural network model was introduced by (El Bazi and Laachfoubi, 2019). The proposed system got two sources of information about words as input: pre-trained word embeddings namely Skip-Gram, CBOW, GLOVE, FastText and Hellinger PCA and character-based representations. This approach yielded an F1-score of 90.60%˙

Table 3 presents the comparison between our models and some state-of-the-art approaches. We outperform the best Arabic NER model (Alsaaran and Alrabiah, 2021) on ANERCorp by about 1.4 points. As far as we know, we are the first to use two types of character-level embedding based on the CNN and LSTM combined with BERT word embeddings as the main feature block for the Arabic NER task. Using these three types of word embeddings together allows the model to capture both important morphological and orthographic patterns as well as the semantic and syntactic word relationships.

6 Conclusion

In this paper, we have presented a technically simple neural network architecture using variant deep neural network architectures and a combinatorial feature embedding based on the CNN, LSTM and BERT. Several deep learning architectures have been investigated and evaluated on ANERCorp dataset. The experimental results have shown that our CLB-BLC and CLB-BGC models achieve superior outcomes over the existing deep learning models in Arabic NER with a high F-score of 93.34% and 93.68% respectively. Also, the evaluation has clearly demonstrated that:

- The CNN model outperforms the LSTM model in extracting character- level features.
- The Bi-GRU model performs better in modeling long-term dependencies for Arabic NER than the Bi-LSTM model. In our experiments, the Bi-GRU architecture outperforms the Bi-LSTM architecture by about 0.34% in the F1-score.
- The proposed models utilizing the combinatorial feature embedding based on the CNN, LSTM and BERT exhibit the highest performance achieving an F1-score of 93.34% and 93.68% F1-score using the BLC and BGC architectures, respectively.

One of the potential directions for future work will be to apply an embedding attention layers on top of the word and character level representations to dynamically determine the information that must be used to best represents the given token and to extend this work to solve the NER task for other languages.

References

Sherief Abdallah, Khaled Shaalan, and Muhammad Shoaib. 2012. Integrating rule-based system with classification for arabic named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 311–322. Springer.

Manel Affi and Chiraz Latiri. 2021. Be-blc: Bert-elmo-based deep neural network architecture for english named entity recognition task. *Procedia Computer Science*, 192:168–181.

Mohammad Al-Smadi, Saad Al-Zboon, Yaser Jararweh, and Patrick Juola. 2020. Transfer learning for arabic named entity recognition with deep neural networks. *Ieee Access*, 8:37736–37745.

Mohammed NA Ali, Guanzheng Tan, and Aamir Hussain. 2018. Bidirectional recurrent neural network approach for arabic named entity recognition. *Future Internet*, 10(12):123.

Mohammed Nadher Abdo Ali, Guanzheng Tan, and Aamir Hussain. 2019. Boosting arabic named-entity recognition with multi-attention layer. *IEEE Access*, 7:46575–46582.

Norah Alsaaran and Maha Alrabiah. 2021. Arabic named entity recognition: A bert-bgru approach. *CMC-COMPUTERS MATERIALS & CONTINUA*, 68(1):471–485.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

David Awad, Caroline Sabty, Mohamed Elmahdy, and Slim Abdennadher. 2018. Arabic name entity recognition using deep learning. In *International Conference on Statistical Language and Speech Processing*, pages 105–116. Springer.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ismail El Bazi and Nabil Laachfoubi. 2019. Arabic named entity recognition using deep learning approach. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(3).

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Palash Goyal, Sumit Pandey, and Karan Jain. 2018. Deep learning for natural language processing. *New York: Apress*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Matthieu Labeau, Kevin Löser, and Alexandre Allauzen. 2015. Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 232–237.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Slim Mesfar. 2007. Named entity recognition for arabic using syntactic grammars. In *International Conference on Application of Natural Language to Information Systems*, pages 305–316. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Chandrahas Mishra and DL Gupta. 2017. Deep machine learning and neural networks: An overview. *IAES International Journal of Artificial Intelligence*, 6(2):66.

Naji F Mohammed and Nazlia Omar. 2012. Arabic named entity recognition using artificial neural network. *Journal of Computer Science*, 8(8):1285.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Mai Oudah and Khaled Shaalan. 2017. Nera 2.0: Improving coverage and performance of rule-based named entity recognition for arabic. *Natural Language Engineering*, 23(3):441–472.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

D A Alzboun Sa'a, Saia Khaled Tawalbeh, Mohammad Al-Smadi, and Yaser Jararweh. 2018. Using bidirectional long short-term memory and conditional random fields for labeling arabic named entities: A comparative study. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 135–140. IEEE.

Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wajdi Zaghouani. 2012. Renar: A rule-based arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1):1–13.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 176–185.

Reputation Analysis Using Key Phrases and Sentiment Scores Extracted from Reviews

Huang Yipu
Graduate School of Science and Engineering, Ibaraki University

Minoru Sasaki
Computer and Information Sciences, Faculty of Engineering, Ibaraki University

Kanako Komiya
Graduate School of Engineering, Tokyo University of Agriculture and Technology

{21nm723s, minoru.sasaki.01}@vc.ibaraki.ac.jp, kkomiya@go.tuat.ac.jp

Abstract

In recent years, reviews of various products by purchasers have been posted on websites. By browsing these reviews, company representatives intend to collect and analyze opinions from the users' side of the products and use them to improve the products. However, existing analysis methods analyze reviews by dividing them into words, and do not analyze the function names that are phrases. Therefore, we propose a reputation analysis method using a decision tree model based on the frequency of key phrases extracted from reviews and sentiment scores. Experimental results showed that the proposed method using key phrases improved the accuracy by 3 points compared to the existing method that analyzes by words.

1 Introduction

When consumers purchase a product or service, they invariably consult reviews. In fact, review analysis has many benefits not only for individuals but also for companies.

For example, the use of the analysis for product planning, analysis of user needs and dissatisfaction, hints for function and service improvement, and verification of the effectiveness of promotions and marketing measures are the effects of corporate review analysis.

In order to improve a product, it is necessary to know what reviewers like about the product and how they evaluate it. Therefore, we use two criteria, the frequency of occurrence of key phrases and the sentiment score, to determine the reviewer's opinion of the product. For example, if the frequency of occurrence of a key phrase is high and the sentiment score is also high, we judge that "everyone is interested in the product, and it has a good reputation" and that the product does not need to be improved immediately.

This time, key phrases (important words) are extracted from a certain review, and a dictionary is created with key phrases as keys and the frequency of occurrence of key phrases as values. Then, we calculate the sentiment score from the review sentences, treat the sentiment score and frequency of key phrases as explanatory variables, and treat whether improvement is necessary or not or priority as objective variables, and create a decision tree model.

The following sections describe key phrase extraction, sentiment score calculation, decision tree model creation, and quality evaluation.

2 Related Research

There have been several studies on review analysis methods. Kobayashi collects and analyzes reviews from web pages, they use TF-IDF to extract keywords, and then extract emotion words, and combine keywords and emotion words (Kobayashi, 2008). However, it is difficult to grasp the whole picture because of the variety of emotion words extracted from various reviews. Abe analyzes product reviews, assigns a score to each evaluation item, and proposes to create an item-by-item dictionary of evaluation expressions related to hotel evaluations, such as keywords, features, and degree (Abe, 2020).

3 Key phrase extraction

3.1 Key Phrase

Key phrase extraction is a technique for extracting phrases that best describe a document. A phrase here is a set of words whose meanings are combined, but in practice, noun phrases are often employed. Many methods already exist, such as PageRank, Text Rank, Single Rank, and Topic Rank (Page, 1999) (Sullivan, 2007).

3.2 Extraction Process

In this case, we will use PKE (Boudin, 2016) to extract key phrases. PKE is produced by Florian Boudin, the author of the MultipartieRank paper.

text	key phrase
バッテリーの方が大きい笑	['バッテリー']
軽量化の為仕方がないのだろう	['軽量化']

Table 1: Examples of Key phrase extraction

Key Phrase Frequency

It is important to calculate the frequency of each key phrase in order to know the customer's level of interest in the product from the reviews (Berger and Mittal, 2000) (Juan-Manuel, 2014). A dictionary is created by adding value values based on the number of times a key phrase appears as shown in Figure 1.

{'バッテリー': 404, '軽量化': 134, '充電器': 501,　　…}

Figure 1: Example of Dictionary

Next is the key phrase frequency calculation, the definition of the key phrase frequency is as follows.

$$z = \frac{x - \mu}{\sigma},$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x - \mu)^2},$$

where x represents the frequency of key phrase and μ represents the Number of key phrases.

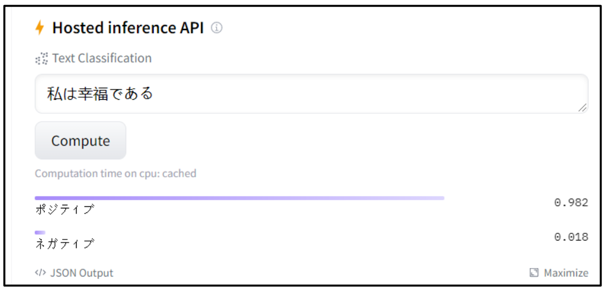
The result of key phrase extraction「トルクは申し分なく握った時の重量配分も問題ない」➞「トルク,申し分,重量配分,問題」➞「1.7452,　-0.923112, 0.4218,　2.9283233」. From this, each review was indicated on the frequency of the key phrase.

4 Emotion Score Calculation

The customer's evaluation is judged to be good or bad based on the review, and the sentiment analysis is used to determine whether the customer's evaluation is good or bad.

4.1 Model

Here we used bert-base-japanese-sentiment from huggingface to evaluate sentiment score as hown in Figure 2.



私は幸福である

Computation time on cpu: cached

ポジティブ 0.982

ネガティブ 0.018

Figure 2: Example of Sentiment Evaluation

4.2 Procedure

In order to effectively perform sentiment analysis here, the sentiment score needs to float between [-1, 1] as shown in Figure 3. In this experiments, Bert-base-Japanese-sentiment model[1] is used to slightly improve the results.

For Positive：

$$score = 2 \times score - 1$$

For Negative：

$$score = 1 - 2 \times score$$

Figure 3: Emotion Scores

4.3 Assignment of Sentiment Scores

One by one, the review sentences are assigned an emotion score. The following are some of the results.

[1] https://huggingface.co/daigo/bert-base-japanese-sentiment (Currently not available)

Text1: そして電池の持ちがいい
Score: 0.9681352376937866
text2: 本体部分が小さいのはびっくりで
score: 0.771661639213562
text3: 多少重いがやむを得ないか
score: -0.9029324054718018
…

5 Decision tree model

5.1 Summary

Decision tree analysis is a data mining method used for “prediction”, “discrimination”, and “classification”. It is an analysis method that finds “explanatory variables” that affect the “dependent variable” of customer information, survey results, etc., and creates a tree-like model.

5.2 Explanatory Variables

The explanatory variables are the objects to be analyzed. Here, the level of interest (frequency of key phrases) and the sentiment score value are done as explanatory variables.

5.3 Objective Variable

The objective variable is the one that displays the results brightly, influenced by the various explanatory variables. Here, the objective variables are set as A, B, C, and D four types as shown in Table 2. The details that each item represents are shown below.

A: Attention and good evaluation
B: Noteworthy and poorly rated
C: Not attracting attention and rated good
D: Not noticed and evaluated poorly

Then, the priority order of improvement for the firms is B D C A. The order of priority for the firms is B D C A.

level of interest	emotion	Objective variable
High	positive	A
	negative	B
low	positive	C
	negative	D

Table 2: Definition of Objective Variables

5.4 Decision Tree Model Construction

If we build a decision tree model, we need explanatory variables and an objective variable. The explanatory variables are the frequency of the key phrases and the sentiment score, previously computed and attached. The objective variable is also an explicitly invisible element. The dataset is divided in a 3:7 ratio between training and test data. The objective variables are then added manually to the training data.

explanatory variable		Objective variable
level of interest	Emotion Score	type
2.923331	0.822314	A
3.22134	-0.92314	B
-0.83167	0.22314	C
1.99913	0.883424	A
…	…	…

Table 3: Decision Tree Model

Now we will analyze the analysis with the CART decision tree (Quinlan, 1987) described in Figure .

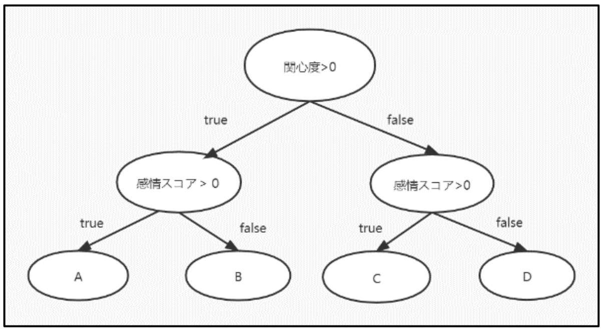


Figure 4 Decision Tree Conceptual Diagram

6 Quality Assessment

6.1 Method

To verify that the model works well, 559 product reviews are used to evaluate the accuracy of the model, which varies with the frequency of key phrases and sentiment scores.

6.2 Results

Using the new 559 reviews, the same key phrase frequency and sentiment scores are transformed, and this model is put in. The results are shown in Table 4. Based on these results, about 83% of the reviews are correct.

	quantity
Results match	468
Difference in results	91

Table 4: Experimental Results

6.3 Analysis and Comparison

The results show that this decision tree model method is effective. Compare the current proposal with the review-analysis methods of Kobayashi et al. and Abe et al. as shown in Table 5.

Our proposed method is a little more accurate than that of Kobayashi et al. because I can grasp the whole picture. Abe et al. used their own evaluation score, but the authenticity of the original evaluation score is still a matter to be examined.

And with our method, we can use the model to clearly show the reviewers and the company's priorities for improvement of the next product. However, there are some areas that need to be improved.

Key word extraction	Emotional Analysis	Model	Accuracy
PKE, Key Phrases	Sentiment score for sentence	Decision Tree Modeling	**83%**
TF-IDF, Keywords	Emotional word extraction	Keyword and Emotional Word	79.9%
Keywords	Review own evaluation score	Target expression dictionary	80.6%

Table 5: Comparison of methods

7 Conclusions

Analysis of reviews is important for companies to improve their products. We propose a CART decision tree with key phrase frequency and sentiment score as explanatory variables to determine the interest and reputation of a product. The four objective variables are used to determine a company's next steps in improving its products. Future work includes improving the key phrase extraction method to increase the accuracy of review analysis and improving the simultaneous decision tree model.

References

Daisuke Kobayashi, Syunichi Nagao, Daisuke Suto, Ushio Inoue. 2008. Proposal of a collection and analysis system of user reviews by text analysis of web pages, DEWS2008 B8-2.

Katsumi Abe and Shinsuke Nakajima. 2020. Hotel recommendation system based on automatic review scoring scheme, DEIM2020 P1-28.

Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web, The Web Conference.

Danny Sullivan. 2007. What Is Google PageRank? A Guide For Searchers & Webmasters, Search Engine Land.

Juan-Manuel Torres-Moreno. 2014. Automatic Text Summarization. John Wiley & Sons.

Adam L. Berger and Vibhu Mittal. 2000. OCELOT: a system for summarizing Web pages, Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Taher H. Haveliwala. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, IEEE Transactions on Knowledge and Data Engineering, 15(4) pp.784–796.

Florian Boudin. 2016. pke: an open source python-based keyphrase extraction toolkit, Proceedings of the 26th International Conference on Computational Linguistics (COLING2016): System Demonstrations, pp.69-73.

John R. Quinlan. 1987. Induction of decision trees. Machine Learning, 1(1) pp.81-106.

Perceptual Overlap in Classification of L2 Vowels: Australian English Vowels Perceived by Experienced Mandarin Listeners

Yizhou Wang
School of Languages and Linguistics
University of Melbourne, Australia
wyz2014tw@gmail.com

Rikke L. Bundgaard-Nielsen
MARCS Institute
University of Western Sydney, Australia
rikkelou@gmail.com

Brett J. Baker
School of Languages and Linguistics
University of Melbourne, Australia
bjbaker@unimelb.edu.au

Olga Maxwell
School of Languages and Linguistics
University of Melbourne, Australia
omaxwell@unimelb.edu.au

Abstract

This paper presents a study examining the role of classification overlap in second language (L2) vowel perception. Eighteen experienced L1 Mandarin Chinese (MC) listeners completed a classification task on twelve Australian English (AusE) monophthongal vowels. It was found that the AusE vowels were not equally difficult for the MC listeners, and the vowels also showed various overlapping patterns in the classification space. Correlation tests revealed that for individual L2 vowels, the classification accuracy was negatively correlated with the mean reaction time (RT), r = -0.817, p = 0.001; the level of classification overlap was negatively correlated with the mean accuracy, r = -0.860, p < 0.001, while positively correlated with the mean RT, r = 0.760, p = 0.004. The findings suggest that classification overlap can inform how experienced L2 listeners' first language (L1) phonology influences vowel perception in a non-native language.

1 Introduction

To successfully understand speech, listeners must process phonetic information and match this information to the phonemic categories of the language in question. For non-native or L2 listeners, this task can sometimes be difficult because their perceptual system is shaped by the L1 phonology. Mismatches between the L1 and L2 phonological systems can lead to perceptual difficulty of various linguistic units in the target language, e.g., segmental (Bundgaard-Nielsen et al., 2011a; Højen and Flege, 2006) and suprasegmental structures (Hallé et al., 2004; Tremblay, 2009). One prevalent theory of non-native speech perception, the Perceptual Assimilation Model (PAM, and its extension, PAM-L2) (Best, 1995; Best and Tyler, 2007) attributes the perceptual difficulty to the interference of the L1 phonological system: L2 phones tend to be assimilated into the listener's L1 phonological categories, which can at times cause a loss of phonemic contrast in perception.

PAM and PAM-L2 posit that the discriminability between a pair of non-native (L2) sounds depends on the level of perceptual overlap between the two sounds, e.g., a pair of perceptually overlapping L2 sounds are more difficult to discriminate than those that do not overlap in cross-language perception (Faris et al., 2018). For instance, L1 Japanese listeners assimilate the AusE vowels [iː] and [ɪə] as their native bimoraic category [iː], while mapping the AusE vowel [ɪ] onto the Japanese monomoraic category [i]; and this perceptual overlap pattern has led them to fail to discriminate between AusE [iː] and [ɪə], though not between AusE [iː] and [ɪ] (Bundgaard-Nielsen et al., 2011a, 2011b).

Since PAM and PAM-L2 provide precise predictions in terms of pairwise discriminability, previous research adopting the framework often uses discrimination tasks (e.g., AXB tasks) to examine perceptual performance by L2 learners. Besides, an assimilation task is used in association to offer predictions of potential sound pairs that show perceptual overlap (Bundgaard-Nielsen et al., 2011a; Faris et al., 2018; Tyler et al., 2014). For a discussion on the methodology, see (Strange and Shafer, 2008). However, discrimination tasks can only examine one pair of L2 sounds at a time, and the assimilation task only offers the native phonemic inventory as the set of classification candidates, which cannot effectively represent the non-native categories that are already learned by *experienced* L2 listeners.

Here, we present a study examining the perception of Australian English (AusE) vowels by experienced L2 listeners whose native language is MC. We use a classification task which allows all target L2 categories to be offered as candidates, and the listener' response is not limited by the range of phonemic categories in their native phonology. In addition, whereas the AXB paradigm only examines one pair of L2 sounds at a time, the classification task used here requires forced phonological processing of the sound stimulus among a set of phonemic categories and thus is able to measure the perceptual similarity between multiple phonemic pairs. These results can potentially reveal scenarios of multiple category assimilation.

To estimate the level of perceptual overlap for a specific pair of L2 sounds, we calculate the *phonological overlap score* following the method reported in Faris et al. (2018). The phonological overlap score of a specific pair of L2 sounds takes into account all classification overlaps that are above chance; and researchers have demonstrated that the score can successfully quantify the perceptual similarity between two sounds (Faris et al., 2016, 2018), especially when the assimilation pattern is not effectively differentiated in the classic framework of PAM (Best, 1995).

The aim of the present study is to directly examine how the perceptual overlap level of L2 vowels links to the overall perceptual difficulty, as indicated by accuracy and latency measures. In particular, the present study aims to answer three research questions:

1. *How are AusE vowels perceived by experienced MC listeners as reflected in the classification results?*
2. *Which AusE vowels show perceptual overlaps when perceived by experienced MC listeners?*
3. *Does the classification performance, accuracy and RT, of AusE vowels correlate with the corresponding perceptual overlap score?*

We hypothesise that, under the influence of native phonological interference, L2 vowel perception can be affected by the level of perceptual overlap between the target L2 sound and other similar categories (possibly more than one). In particular, we hypothesise that the perceptual accuracy should be negatively correlated with the perceptual overlap level; and that the RT measure should be positively correlated with the overlap level. This prediction is also consistent with the Automatic Selective Perception (ASP) model (Strange, 2011; Strange and Shafer, 2008) which posits that difficult L2 sounds tend to take longer to process than easy categories.

2 Methods

2.1 Participants

Eighteen L1 MC listeners participated in the study. The participants were first-year international students at an Australian university (M_{age} = 23.6, range: 21–27). All spoke English as a second language, and their mean length of English education was 16.2 yr (*SD* = 3.6, range: 10–24). None reported a history of speech or hearing disorders. All participants were experienced L2 learners with advanced target language proficiency (B2-C1 level), and they were able to complete the perception task directly in the target language, i.e., English.

2.2 Stimuli

All twelve AusE monophthongal vowels [iː ɪ e eː æ ɐ ɐː ɜː ɔ oː ʊ ʉː] (as in *beat*, *bit*, *bet*, *paired*, *bat*, *but*, *Bart*, *bird*, *bot*, *bought*, *boot*, and *who'd*) were included in the study (excluding only schwa). All the vowel stimuli were generated in the phonetic software Praat (Boersma, 2001), and the values for F1, F2, and duration were the same as those reported in (Cox, 2006) for AusE male speakers. The pitch contour dropped linearly from 120 Hz to 100 Hz for all vowels, and the intensity was also set identically at a level of 70 dB SPL for all the stimuli. These manipulations ensured that listeners only attended to the first two formants and the characteristic duration value when making classification decisions.

2.3 Procedure and analysis

All participants completed a vowel classification task in the phonetic software Praat (Boersma, 2001) on a laptop computer. Classification responses were made by clicking boxes with keyword labels. Stimuli were delivered via acoustic headphones, and both the responses and the RTs were automatically recorded. Each participant completed 576 trials (12 vowels, 48 repetitions), and the stimuli were presented in a pseudo-randomised order. An additional 60 trials were given in a practice block before the experiment for familiarisation. To address research question (1), classification results were aggregated to generate a 12 × 12 confusion matrix in which each cell represents a unique "stimulus-response" combination. Any cell with a response percentage over 70% is considered as a "categorised" case, which indicates a consistent perceptual pattern (Bundgaard-Nielsen et al., 2011a, 2011b; Tyler et al., 2014). Classification is deemed as correct if the response category matches the stimulus phoneme. To address the research question (2), we calculated the between-category overlap score following the method introduced by Faris et al. (2018), which is formulated as below:

$$\boxed{Overlap(AB) = \sum_{1}^{n} min\{P_n(A), P_n(B)\}}$$

where *A* and *B* represent two L2 categories, *n* represent the number of above chance response categories, *Pn(A)* and *Pn(B)* represent the percentage of choice of a specific vowel as the corresponding above chance category, and the function *min{}* selects the smaller value of the two percentage scores. In the present study, since there are twelve alternatives, the chance level is 1/12 ≈ 8%.

To address research question (3), the between-category perceptual overlap scores were further analysed on the basis of every single AusE vowel. For instance, some vowels can be overlapping with multiple categories, while in optimal performance, no perceptual overlap should be observed if every vowel is correctly classified in all trials. For each vowel, a sum score of perceptual overlap can be added up from all the pairwise overlaps where the target vowel is involved. To confirm the hypotheses, a negative correlation is expected between the classification accuracy and the sum score of perceptual overlap, and a positive correlation is expected between the perceptual overlap and RT.

3 Results

3.1 Classification results

The overall results of the classification task are summarised in Table 1. Cells on the diagonal line represent the average categorisation accuracy of each AusE vowel. The AusE vowel [ɪ] reached the highest accuracy of 90%, with an average RT of 1.60 sec, while [æ] was associated with the lowest accuracy of 41%, and the average RT was 2.98 sec. When the 70% criterion was applied (Bundgaard-Nielsen et al., 2011a, 2011b), seven AusE vowels were considered as categorised: [ʉː], [ɜː], [ɐː], [iː], [ʊ], [ɔ], and [ɪ]; while the other five vowels fail to reach the criterion: [æ], [eː], [e], [oː], and [ɐ]. A general tendency was observed that the vowels with higher classification accuracy values have shorter RT measures. To confirm this tendency, a Pearson's correlation test revealed that the two measures were negatively correlated, *r* = -0.817 (95%CI: -0.459 ~ -0.947), *p* = 0.001.

3.2 Pairwise perceptual overlap

In the present study, we tested the classification of twelve AusE vowels, which potentially derive $C_{(12,2)}$ = 66 pairwise contrasts. However, not all of the vowel pairs were perceived as similar or overlapping by the MC listeners. Following previous research (Faris et al., 2018), we calculated the phonological overlap scores for all the 66 potential contrasts and find fifteen pairs that have shown above chance (> 8%) overlap patterns, see Table 2. The AusE vowel pair [æ]-[e] received the highest overlap score of 44, followed by the [æ]-[eː] pair with an overlap score of 34, [oː]-[ʊ] with a

Table 1. *Classification matrix of AusE vowels, percentage (mean RT, in seconds)*

	Response Category											
Stim.	bad [æ]	paired [eː]	bed [e]	bought [oː]	but [ɐ]	who'd [ʉː]	bird [ɜː]	Bart [ɐː]	beat [iː]	hood [ʊ]	bot [ɔ]	bit [ɪ]
[æ]	41 (2.98)	8 (3.67)	27 (3.36)		5 (3.46)		17 (3.16)	1 (4.54)				
[eː]	4 (4.67)	55 (2.96)	10 (4.06)			7 (4.02)	16 (3.29)		4 (2.62)	1 (3.44)		2 (2.88)
[e]	9 (3.86)	15 (3.53)	56 (3.15)			4 (3.19)	5 (3.26)					10 (2.14)
[oː]				66 (2.27)		8 (3.18)				15 (3.34)	11 (3.03)	
[ɐ]	7 (3.73)	1 (4.26)	3 (4.05)		67 (2.48)		7 (3.39)	14 (2.93)				
[ʉː]		4 (6.07)	4 (4.71)			**71** **(2.68)**	8 (3.29)			11 (3.46)		
[ɜː]	2 (5.54)	8 (4.31)	5 (5.27)		1 (3.82)	10 (2.96)	**72** **(2.94)**	1 (4.87)				
[ɐː]	2 (5.81)	2 (3.98)			7 (4.70)		8 (4.13)	**81** **(2.26)**				
[iː]		3 (3.40)				5 (2.71)			**81** **(1.86)**			9 (2.35)
[ʊ]				1 (2.27)		7 (3.57)				**81** **(2.40)**	11 (2.26)	
[ɔ]				10 (3.51)	3 (2.27)		2 (3.76)			2 (3.63)	**82** **(1.90)**	
[ɪ]			1 (2.99)			1 (4.34)			4 (2.76)	4 (1.95)		**90** **(1.60)**

Note. Cells with a percentage under 1% were removed; identification percentages over 70% are in boldface.

score of 25, [eː]-[e] with a score of 25, and [oː]-[ɔ] with a score of 20. Other overlapping pairs had a score under 20.

Table 2. *Pairwise perceptual overlap scores*

No.	AusE Vowel	Overlap
1	[æ]-[e]	44
2	[æ]-[eː]	34
3	[oː]-[ʊ]	25
4	[eː]-[e]	25
5	[oː]-[ɔ]	20
6	[æ]-[ɜː]	17
7	[eː]-[ɜː]	16
8	[ɐ]-[ɐː]	14
9	[oː]-[ʉː]	11
10	[ʉː]-[ʊ]	11
11	[ʊ]-[ɔ]	11
12	[ʉː]-[ɜː]	10
13	[e]-[ɪ]	10
14	[e]-[iː]	9
15	[iː]-[ɪ]	9

According to PAM and PAM-L2 (Best, 1995; Best and Tyler, 2007; Faris et al., 2018), the perceptual similarity between L2 vowels can predict their mutual discriminability. The present study has found that perceptual overlap of vowels seems to be gradient, e.g., both the mid vowels [e] and [ɜː] were perceived as overlapping with [æ] and [eː], but to different degrees: [e]-[æ] (44) > [ɜː]-[æ] (17), and [e]-[eː] (25) > [ɜː]-[eː] (16). According to PAM, high levels of perceptual overlap lead to the assimilation types that are difficult to discriminate, including single-category (SC type), category-goodness (CG type), and some uncategorised-uncategorised (UU type, with classification overlaps) assimilations (Best, 1995; Best and Tyler, 2007; Faris et al., 2018).

3.3 Classification accuracy and overlap

The previous section has reported that all twelve AusE vowels are perceived as overlapping with at least one other category, and the level of perceptual overlap also differs from pair to pair. In addition, some vowels share overlaps with more than one other vowel category, e.g., the low front vowel [æ] was perceived as overlapping with [e] (44), [eː] (34), and [ɜː] (17), which yields a sum score of 95. This calculation probes into how much the overlapping area is taken up in the full classification space, and also estimate the likelihood of multiple category assimilation. Following this method, the sum scores for the other eleven AusE vowels were also calculated, and the values are summarised in Table 3, together with their mean accuracy and RT measures.

Table 3. *Overlap sum score and classification accuracy*

AusE Vowel	Overlap (Sum)	Accuracy (%)	RT (Sec)
[æ]	95	41	2.98
[e]	88	56	3.15
[eː]	74	55	2.96
[oː]	56	66	2.27
[ʊ]	47	81	2.40
[ɜː]	43	72	2.94
[ʉː]	32	71	2.68
[ɔ]	31	82	1.90
[iː]	19	81	1.86
[ɐ]	14	67	2.48
[ɐː]	14	81	2.26
[ɪ]	10	90	1.60

The analysis showed that three front vowels [æ], [e], and [eː] receive the highest sums of 95, 88, and 74, respectively. Note that these three vowels also showed the lowest accuracy measures of 41%, 56%, and 55%, respectively. These responses were also made with the highest level of decision uncertainty, as indicated by the RT measure of 2.98, 3.15, and 2.96 sec. The short high front vowel [ɪ] had the lowest total score of perceptual overlap (10), and it was also the category that showed the highest accuracy (90%). To confirm this tendency and answer research question (iii), a Pearson's correlation test was carried out, and it confirmed that the sum score of perceptual overlap was negatively correlated with the mean accuracy for the twelve AusE vowels, r = -0.860, 95% CI: -0.565 ~ -0.960, p < 0.001. For the RT measure, it was found to be positively correlated with the sum overlap score, r = 0.760, 95%CI: 0.330 ~ 0.929, p = 0.004. These results confirmed our hypotheses that both the classification accuracy and speed are closely associated with the perceptual overlap level.

4 Discussions

In the present study, we have examined the perception of the whole monophthong system of AusE by experienced L2 listeners whose native language is MC. In answering research question (i), we analysed the relative difficulty of vowel perception using both accuracy and latency measures. The findings suggest that not all the AusE vowels are easy to perceive even after more than ten years of English study. The negative correlation between classification accuracy and RT found in our data is consistent with the Automatic Selective Perception (ASP) model (Strange, 2011; Strange and Shafer, 2008), which predicts that successfully learned sound categories are perceived in a more automatic manner by non-native (L2) listeners. In answering research question (ii), we have diagnosed fifteen AusE vowel pairs that may be vulnerable to category merging and discrimination problems. In particular, our findings of the perceptual similarity patterns are consistent with reports of pairwise discrimination studies based on other English dialects, e.g., Canadian English (Wang, 2006; Wang and Munro, 1999, 2004), and American English (Jia et al., 2006; Lai, 2010).

In answering research question (iii), we extend the notion of perceptual overlap from pairwise discrimination towards an application in a multiple-choice classification task. The overlap score method used in the present study was introduced by a perceptual assimilation study which aimed to differentiate the subtypes of uncategorised assimilations (Faris et al., 2018). While the classic framework of PAM and PAM-L2 makes predictions based on qualitative differences in terms of assimilation mapping, e.g., two-category assimilation type means two L2 sounds are categorised as two distinct native categories, and excellent discrimination is expected (Best, 1995; Best and Tyler, 2007). The overlap score method (Faris et al., 2018) enables us to describe the perceptual similarity in quantitative terms.

The importance of analysing misclassification patterns is often neglected in previous studies, but for L2 phoneme classification the error patterns are not purely random: certain vowels are more likely to be confused than others, e.g., MC listeners can confuse between AusE [æ] and [e], but less often between [æ] and [ɐ], see Table 1, 2.

Here we offer a schematic explanation of perception overlap in a minimal scenario, where A, B, and C are three non-native (L2) phonemes, and X and Y are two native (L1) phonemes: if both A and B are both perfectly mapped to category X (i.e., single-category assimilation), and C is mapped to category Y, then an unbiased listener should be unable to distinguish A from B, and should thus randomly categorise the sounds using the two labels, which will cause the classification results to be highly overlapping, see Figure 1. On the other hand, category C will show non-overlapping classification patterns with either A or B, because C forms a two-category assimilation pair with them. To conclude, classification overlaps reflect the level of perceptual similarity between L2 phonemes, and the patterns also reveal how these non-native sounds are assimilated by the native phonological system. As for experienced L2 listeners, the perceptual overlap between categories is more likely to be partial rather than a complete merging, because new L2 categories can be developed through extensive phonetic exposure and learning.

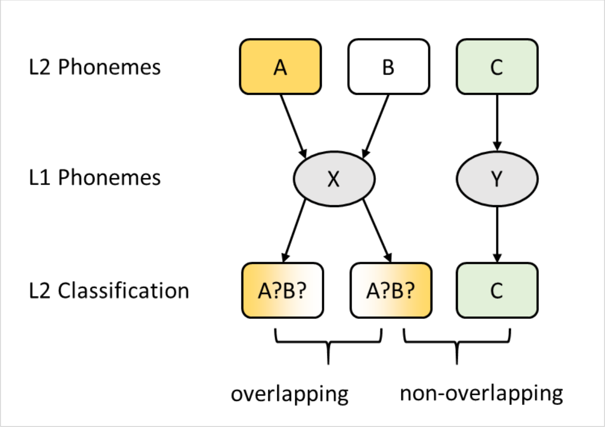


Figure 1. The relation between perceptual assimilation and L2 classification.

The present study, however, does have some limitations. Since the vowel stimuli used in the present classification task are presented in isolation, we have not examined whether different phonological contexts can influence the L2 listeners' classification performance, while a growing body of research has reported that vowel perception is sensitive to the phonetic-phonological context in which the target vowel is situated (Levy, 2009a, 2009b; Strange et al., 2001). Additionally, since the vowel stimuli were synthesised signals with invariant formant values, the dynamic nuances of monophthongal vowels are ignored in the present study, while some studies have found that authentically produced vowel dynamics such as the inherent spectral change are important cues in vowel recognition (Jin and Liu, 2013; Morrison, 2009). Finally, since our study uses the classification paradigm, which is not often used by perceptual assimilation studies, further experiment comparison research is needed to inform the correspondence between the patterns revealed by a classification task and those from more conventional paradigms, e.g., assimilation and discrimination tasks (Bundgaard-Nielsen et al., 2011a; Faris et al., 2018; Tyler et al., 2014). In particular, how does classification overlap score (as shown in Table 2) corresponds to mutual discriminability remains an urgent topic for future studies.

5 Conclusion

In the present study, we have used the perceptual overlap score to analyse confusion patterns in L2 vowel perception, and we have also compared the overlap scores with the accuracy and RT measures and found robust correlation properties between them. These findings suggest that the misclassification patterns in a confusion matrix can reveal the perceived similarity between sound categories. For L2 listeners, in particular, this information should be analysed together with the potential interference from the listeners' native phonology, e.g., perceptual assimilation. The method used in the present study has the potential to be incorporated with other experimental paradigms (e.g., assimilation and discrimination tasks) in the cases where L2 listeners have relatively high proficiency in the target language and therefore have access to the emerging L2 categories.

References

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn (Ed.), *Language experience in second language speech perception* (pp. 13–34). John Benjamins.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9–10), 341–345.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011a). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *22*, 433–461. https://doi.org/10.1017/S0272263111000040

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011b). Vocabulary size matters: The assimilation of second language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics*, *32*(1), 51–67. https://doi.org/10.1017/S0142716410000287

Cox, F. (2006). The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics*, *26*(2), 147–179.

Faris, M. M., Best, C. T., & Tyler, M. D. (2016). An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *The Journal of the Acoustical Society of America*, *139*(1), EL1–EL5. https://doi.org/10.1121/1.4939608

Faris, M. M., Best, C. T., & Tyler, M. D. (2018). Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories. *Journal of Phonetics*, *70*, 1–19. https://doi.org/10.1016/j.wocn.2018.05.003

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. https://doi.org/10.1016/S0095-4470(03)00016-0

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America*, *109*, 3072–3084.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118–1130.

Jin, S.-H., & Liu, C. (2013). The vowel inherent spectral change of English vowels spoken by native and non-native speakers. *The Journal of the Acoustical Society of America*, *133*(5), EL363–EL369. https://doi.org/10.1121/1.4798620

Lai, Y. (2010). English vowel discrimination and assimilation by Chinese-speaking learners of English. *Concentric: Studies in Linguistics*, *36*(2), 157–182.

Levy, E. S. (2009a). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America*, *125*(2), 1138–1152. https://doi.org/10.1121/1.3050256

Levy, E. S. (2009b). On the assimilation-discrimination relationship in American English adults' French vowel learning. *The Journal of the Acoustical Society of America*, *126*(5), 2670–2682. https://doi.org/10.1121/1.3224715

Morrison, G. S. (2009). L1-Spanish Speakers' Acquisition of the English /i-ɪ/ contrast II: Perception of Vowel Inherent Spectral Change. *Language and Speech*, *54*(4), 437–462.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America*, *109*(4), 1691–1704.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners. In J. G. Hansen-Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 153–192). Benjamins.

Tremblay, A. (2009). Phonetic variability and the variable perception of L2 word stress by french canadian listeners. *International Journal of Bilingualism*, *13*(1), 35–62. https://doi.org/10.1177/1367006909103528

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica*, *71*(1), 4–21. https://doi.org/10.1159/000356237

Wang, X. (2006). Mandarin listeners' perception of english vowels: Problems and strategies.

Canadian Acoustics, *34*(4), 15–26.

Wang, X., & Munro, M. J. (1999). The perception of English tense-lax vowel pairs by native Mandarin speakers: The effects of training on attention to temporal and spectral cues. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (Vol. 3, pp. 125–128). American Institute of Physics.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539–552.

Sinhala Sentence Embedding: A Two-Tiered Structure for Low-Resource Languages

Gihan Weeraprameshwara[*,1], **Vihanga Jayawickrama**[*], **Nisansa de Silva**[*], and **Yudhanjaya Wijeratne**[**]

[*]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[**]LIRNEasia, Sri Lanka
[1]gihanravindu.17@cse.mrt.ac.lk

Abstract

In the process of numerically modeling natural languages, developing language embeddings is a vital step. However, it is challenging to develop functional embeddings for resource-poor languages such as Sinhala, for which sufficiently large corpora, effective language parsers, and any other required resources are difficult to find. In such conditions, the exploitation of existing models to come up with an efficacious embedding methodology to numerically represent text could be quite fruitful. This paper explores the effectivity of several one-tiered and two-tiered embedding architectures in representing Sinhala text in the sentiment analysis domain. With our findings, the two-tiered embedding architecture where the lower-tier consists of a word embedding and the upper-tier consists of a sentence embedding has been proven to perform better than one-tier word embeddings, by achieving a maximum F1 score of 88.04% in contrast to the 83.76% achieved by word embedding models. Furthermore, embeddings in the hyperbolic space are also developed and compared with Euclidean embeddings in terms of performance. A sentiment data set consisting of Facebook posts and associated reactions have been used for this research. To effectively compare the performance of different embedding systems, the same deep neural network structure has been trained on sentiment data with each of the embedding systems used to encode the text associated.

1 Introduction

An effective numerical representation of the textual content is crucial for natural language processing models, in order to understand the underlying relational patterns among words and discover patterns in natural languages. For resource-rich languages like English, numerous pre-trained models as well as the required materials to develop an embedding system are readily available. On the contrary, for resource-poor languages such as Sinhala, neither of those options could be easily found (de Silva, 2019). Even the data sets that are available for training often fail to meet adequate standards (Caswell et al., 2021). Thus, discovering a convenient methodology to develop embeddings for text would be a great step forward in the NLP domain for the Sinhala language.

Sinhala, also known as Sinhalese, is an Indo-Aryan language that is used within Sri Lanka (Kanduboda, 2011). The primary user base of this language is the Sinhalese ethnic group of the country. In total, 17 million people use Sinhala as their first language while 2 million people use it as a second language (de Silva, 2019). Furthermore, Sinhala is structurally different from English, which uses a subject-verb-object structure as opposed to the subject-object-verb structure used by Sinhala as shown in the figure 1 thus most of the pre-trained embedding models for English may not be effective with Sinhala.

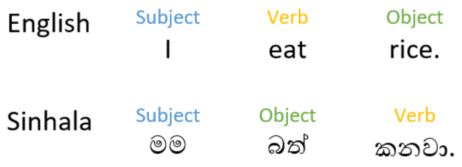


Figure 1: SVO grammar structure of English and SOV grammar structure of Sinhala

This study therefore is focused on discovering an effective embedding system for Sinhala text that provides reasonable results when used in training deep learning models. Sentiment analysis with Facebook data is utilized as the use case for the study.

Upon considering common forms of vector presentations of textual content, bag of words, word embedding, and sentence embedding are three of the leading methodologies in the present. Word embeddings have been observed to surpass the performance of bag of words for large enough data sets (Rudkowsky et al., 2018) because bag of words often met with various problems such as disregarding the grammatical structure of the text, large vocabulary dimension and sparse representation (Le and Mikolov, 2014; El-Din, 2016). In order to tackle the above challenges, word embeddings can be used. Since word embeddings capture the similarities among ingrained sentiments in words and represent them in the vector space, word embeddings tend to increase the accuracy of classification models (Goldberg, 2016).

However, one of the major weaknesses of word embedding models is that they fail to capture syntax and polysemy; i.e. the presence of multiple possible meanings for a certain word or a phrase (Mu et al., 2016). In order to overcome these obstacles and also to achieve fine granularity in the embedding, sentence embeddings are used. The idea is to test common Euclidean space word embedding techniques such as fastText (Bojanowski et al., 2017; Joulin et al., 2016), Word2vec (Mikolov et al., 2013), and GloVe (Pennington et al., 2014) with sentence embedding techniques. The pooling methods (i.e. max pooling, min pooling and avg pooling) will be considered as the baseline methods for the test. More advanced models such as sequence to sequence model (i.e. seq2seq model) (Sutskever et al., 2014) and the modified version of the sequence to sequence model introduced by the work of Cho et al. (2014) with GRU (Chung et al., 2014) and LSTM (Hochreiter and Schmidhuber, 1997) recurrent neural network units will be tested against the pooling means. Furthermore, the addition of attention mechanism (Vaswani et al., 2017) into the sequence to sequence model will also be tested.

Most models created using word and sentence embeddings are based on the Euclidean space. Though this vector space is commonly used, it poses significant limitations when representing complex structures (Nickel and Kiela, 2017). Using the hyperbolic space provides a plausible solution for such instances. The hyperbolic space is a negatively-curved, non-Euclidean space. It is advantageous for embedding trees as the circumference of a circle grows exponentially with the radius. The usage of hyperbolic embedding is still a novel research area as it was only introduced recently, through the work of Nickel and Kiela (2017); Chamberlain et al. (2017); Sala et al. (2018). The work of Lu et al. (2019, 2020) highlight the importance of using the hyperbolic space to improve the quality of embeddings in a practical context within the medical domain. However, research done on the applicability of hyperbolic embeddings in different arenas is highly limited. Thus, the full potential of the hyperbolic space is yet to be fully uncovered.

Through this paper, we are testing the effectiveness of a set of two-tiered word representation models that include various word embeddings as the lower tier and sentence embeddings as the upper tier will be compared.

2 Related Work

The sequence to sequence model introduced by the work of Sutskever et al. (2014) is vital in this research as it is one of the core models in developing sentence embedding. Though originally developed for translation purposes the model has gone under multiple modifications depending on the context such as description generation for images (Karpathy and Fei-Fei, 2015), phrase representation (Cho et al., 2014), attention models (Vaswani et al., 2017) and BERT models (Devlin et al., 2018) thus proving the potential it holds in the machine learning area.

The work of Nickel and Kiela (2017) introduces and explores the potential of hyperbolic embedding by using an n-dimension Poincaré ball. The research work compares the hyperbolic and Euclidean embeddings for a complex latent data structure and comes to the conclusion that hyperbolic embedding surpasses the Euclidean embedding in effectivity. Inspired by the above results, both Leimeister and Wilson (2018) and Dhingra et al. (2018) have extended the methodology introduced by Nickel and Kiela (2017). Leimeister and Wilson (2018) have developed a hyperbolic word embedding using the skip-ngram negative sampling architecture taken from Word2vec. In lower embedding dimensions, the developed model performs better in comparison to its Euclidean counterpart. The work of Dhingra et al. (2018) uses re-parameterization to extend the Poincaré embedding, in order to learn the embedding of arbitrarily parameterized objects. The framework thus created is used to develop word and sentence embeddings. In our research, we will be following the footsteps of the above papers.

When considering the usage of hyperbolic embeddings in a practical context, the work of Lu et al. (2019, 2020) can be examined. The research by Lu et al. (2019) improves the state-of-the-art model used to predict ICU (intensive care unit) re-admissions and surpasses the accepted benchmark used to predict in-hospital mortality using hyperbolic embedding of Electronic Health Records, while the work of Lu et al. (2020) introduces a novel network embedding method which is capable of maintaining the consistency of the node representation across two views of networks, thus emphasizing the capabilities of hyperbolic embeddings. To the best of our knowledge, hyperbolic embeddings have not been previously applied to Sinhala content. Therefore, this research may reveal novel insight regarding hyperbolic embedding and its effectivity in sentiment analysis.

In the research work of Senevirathne et al. (2020), capsule-B model (Zhao et al., 2018) is crowned as the state-of-the-art model for the Sinhala sentiment analysis. In this work, a set of deep learning models are tested for the ability to predict the sentiment of Sinhala news comments. The GRU (Chung et al., 2014) model with a CNN (Wang et al., 2016) layer which is used for the testing of each embedding in this work is taken from the aforementioned research. Furthermore, the work of Weeraprameshwara et al. (2022) has extended the idea and tested the same set of deep learning models with the addition of sentiment analysis models introduced in the work of Jayawickrama et al. (2021) using the Facebook data set which is used in this research work. According to their results, the 3 layer stacked BiLSTM model (Zhou et al., 2019) outshines as the state-of-the-art model.

3 Methodology

In order to test the feasibility of two-tiered word representation as a means of representing Sinhala text in the sentiment analysis domain, a series of experiments were conducted as described in the following subsections.

3.1 Data Set

The data set used for the project is extracted from the work of Wijeratne and de Silva (2020), which contains 1,820,930 Facebook posts from 533 Facebook pages popular in Sri Lanka over the time window of 2010 to 2020. The research work has produced two cleaned corpora and a set of stop words for the given context. The larger corpus among them consists of a total of 28 to 29 million words. The data set covers a wide range of subjects such as politics, media, and celebrities. Table 1 illustrates the fields taken from the data set for the embedding development, model training and testing phases.

Field Name	Total Count	Percentage(%)
Likes	312,282,979	93.58
Loves	10,637,722	3.19
Wow	1,633,255	0.49
Haha	5,377,815	1.61
Sad	2,611,908	0.78
Angry	1,158,182	0.35
Thankful	12,933	0.00

Table 1: The counts and percentages of the reactions in the Facebook data set

3.2 Preprocessing

Even though there are two preprocessed corpora introduced through the work of Wijeratne and de Silva (2020), the raw data set was used for this research with the objective of preprocessing it to suit our requirements. As such, numerical content, URLs, email addresses, hashtags, words in other languages except for Sinhala and English, and excessive spaces were removed from the text. While the focus of this study is colloquial Sinhala, English is included in the data set as the two languages are often codemixed in colloquial use. Codemixing of Sinhala with other languages is much less in comparison. Furthermore, stop words were removed from the text as well, as recommended by Wijeratne and de Silva (2020). Posts with no textual content after thus preprocessing as well as posts with no reaction annotations were also removed as they yield no value in the annotation stage. The final preprocessed data set consists of Sinhala, English, and Sinhala-English code mixed content, adding up to a total of 542,871 Facebook posts consisting of 8,605,849 words.

3.3 Annotation

Since the procedure followed in the model development is supervised learning, the data set needed to be annotated (Schapire and Freund, 2012). It is quite a considerable challenge to obtain sufficiently large annotated data sets for resource-poor languages like Sinhala thus Facebook data set is ideal for the given scenario as the Facebook posts are pre-annotated by Facebook users using Facebook reactions. Though this is not an expert annotation, it can be considered as an effective means of community annotation as the collective opinion of a large number of Facebook users is represented by the reaction annotation (Pool and Nissim, 2016; Freeman et al., 2020; Graziani et al., 2019; Jayawickrama et al., 2021).

A binary classification method which was introduced through the work of Senevirathne et al. (2020) and further improved for Facebook data by Weerapremeshwara et al. (2022) is used in this research as the annotation schema which is illustrated in the figure 2. Here, the Facebook reactions are divided into two classes; positive reactions and negative reactions. The reactions *love* and *wow* are considered as positive reactions while *sad* and *angry* are classified as negative reactions. The reactions *like* and *thankful* have been excluded as they are outliers in the data set with respect to the other reactions. The *like* is the de facto reaction given by the users and it does not yield a valid sentiment. The *thankful* reaction has appeared in a small time period making the presence insignificant compared to other reactions (only 0.00003% of the total reaction count). The *haha* reaction is also excluded due to the contradicting nature of its use cases (Jayawickrama et al., 2021). The *care* reaction is not included in this data set as it was first introduced to the platform in 2020 (Lyles, 2020), after the creation of the data set.

3.4 Word Embeddings

The final vector representation of Facebook posts consists of two major elements: word embeddings and sentence embeddings.

Word embeddings are used both as the first tier of the two-tiered embedding systems and as the basic one-tiered embedding systems used in the form of a benchmark against which the performance of two-tiered embedding systems would be compared. The performance of both Euclidean and hyperbolic word embeddings has been thus evaluated in this research.

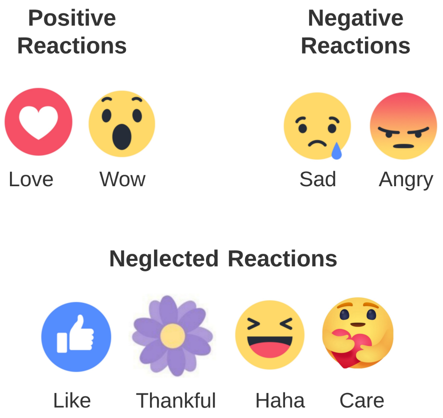


Figure 2: Reaction categorization for the annotation

3.4.1 Euclidean Word Embeddings

For the purpose of representing words in the Euclidean space; fastText, Word2vec, and GloVe word embedding techniques were utilized. Word vectors consisting of 200 dimensions were created using each of the aforementioned models and a window size of 40 was picked based on the work of Senevirathne et al. (2020); Weerapremeshwara et al. (2022) which precedes this research.

3.4.2 Hyperbolic Embeddings

The hyperbolic space exhibits different mathematical properties in comparison to the Euclidean space. Due to its inherent properties, the Euclidean space struggles to model a latent hierarchy. This issue could be addressed by mapping the embedding into a higher dimension (Nickel and Kiela, 2017). However, this may lead to sparse data mapping, causing the curse of dimensionality to affect the performance. This may induce adverse effects such as causing the machine learning model to overfit by the data and using a high memory capacity for computations and storage.

The hyperbolic space has caught the attention of researchers as a plausible solution to such issues encountered in using the Euclidean space for modeling complex structures. The unique feature of this mathematical model is that the space covered by an n-ball in an n-dimensional hyperbolic space increases exponentially with the radius. In contrast to the Euclidean space where the space covered by an n-ball remains restricted by the n^{th} power of the radius, the hyperbolic space could easily handle complex models such as tree-like structures within a limited dimensionality.

The distance (D) between two vectors (i and j) in the hyperbolic space can be calculated as shown in equation 1.

$$D_{(i,j)} = \text{arcCosh}\,(1 + \frac{2||i-j||^2}{(1-||i||^2)(1-||j||^2)}) \tag{1}$$

Gaussian curvature is denoted by K. The circumference (C) of a hyperbolic circle with radius r is calculated as displayed in equation 2 while the area (A) can be calculated using equation 3.

$$C = 2\pi R \sinh{(r/R)} \tag{2}$$

$$A = 2\pi R^2 \cosh{(r/R-1)} \tag{3}$$

Here, R is a constant of which the value is depicted by equation 4.

$$R = \frac{1}{\sqrt{-K}} \tag{4}$$

Since both the circumference and the area of a hyperbolic circle grow exponentially with the radius, the hyperbolic space has the capability to effectively store a complex latent hierarchy of data using a much lower number of dimensions than the Euclidean space would require to store the exact same structure.

In order to create hyperbolic word embeddings, the data set should be reformed in such a manner that the syntactic structure of data is highlighted. However, an adequate language parser for Sinhala does not currently exist (de Silva, 2019). Using parsers dedicated to the English language is also unfitting since the underlying grammatical structure of Sinhala is significantly different from that of English. Furthermore, for codemixed colloquial data present in this data set, grammatical structures of both Sinhala and English languages would have to be taken into consideration. Therefore, the parsing mechanism shown in figure3 is used to generate word tokens. A total of 8605849 tokens have been thus generated.

The two-dimensional illustration of the Poincaré ball after training with the Facebook data set is shown in the figure 4. Each node represents a word in the figure and each edge represents the connection between words. Here for the illustration purposes, only a thousand nodes are shown and the dimension is projected from 200 to 2.

The clustering of semantically related words in the Poincaré embedding is shown in the figure 5.

Sentence :- මම bus එකේ ගෙදර යනවා

Relations :- {මම: මම bus එකේ ගෙදර යනවා}
{bus: මම bus එකේ ගෙදර යනවා}
{එකේ: මම bus එකේ ගෙදර යනවා}
{ගෙදර: මම bus එකේ ගෙදර යනවා}
{යනවා: මම bus එකේ ගෙදර යනවා}

Figure 3: Examples of parsing mechanism used for the hyperbolic embeddings where each word is matched to the sentence

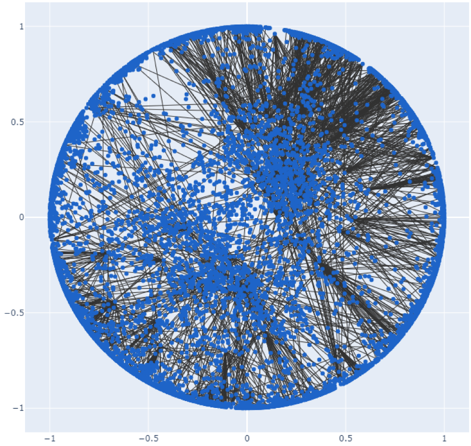


Figure 4: Poincaré word embedding done on Facebook data set

A set of words related to cricket sport is clustered in the top left corner while a set of Sinhala words related to Christianity is clustered in the bottom left. A cluster which represents news-related terms is formed in the bottom right corner. With this evidence, we can safely assume that the hyperbolic space has the capability to store a complex latent hierarchy such as the semantic relation of words.

3.5 Sentence Embeddings

Sentence embeddings are used as the second tier of the two-tiered embedding models. Basic pooling methods as well as the sequence to sequence model are used to generate sentence embeddings by using the word embedding of each word in a sentence. For both Euclidean space and hyperbolic space embeddings, the sentence embeddings are generated in a similar fashion as described in section 3.5.1.

With the parsing method mentioned in the section 3.4.2 the sentence embeddings of the Poincaré embedding is formed as illustrated in the figure 6. Both sentences and words included in them are added to the Poincare ball in the form of nodes, with nodes representing words surrounding those

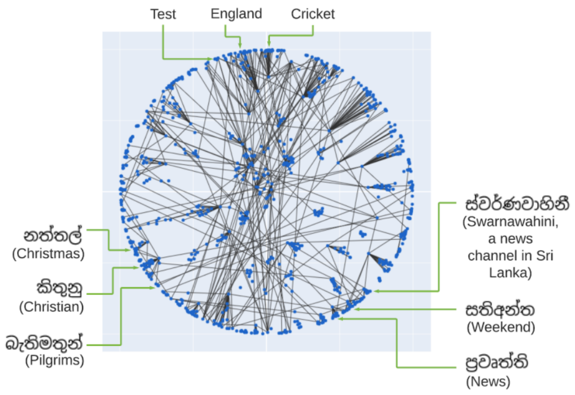


Figure 5: Word clustering in the Poincaré embedding. The meaning of the Sinhala words are given in the brackets

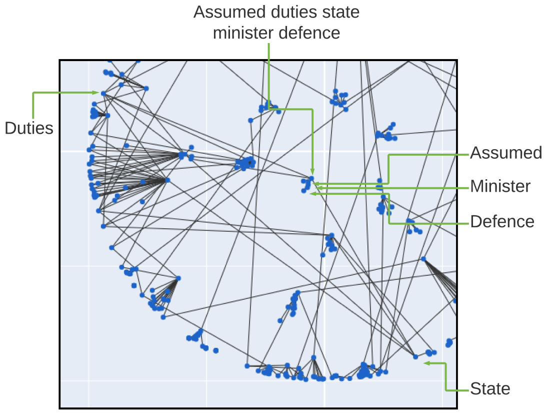


Figure 6: The sentence vector representation in the Poincaré embedding

representing the sentences they are used in. Words that are used more frequently in the data set are pushed further towards the edge of the ball while less frequent words reside closer to the centre. The vector representing a sentence stripped of stop words would be placed closer to the centre of the ball as well, since a sentence would be repeated much less frequently in the data set in comparison to a word. As figure 6 portrays, the word *defence* which is not frequently used in the data set is located further away from the edge of the ball than the word *state*, which is used more frequently.

3.5.1 Pooling

Sentence embeddings have been created with three different pooling mechanisms for each of fastText, Word2vec, GloVe, and hyperbolic word embeddings; namely, max pooling, min pooling, and avg pooling. Pooling embeddings will be considered as baseline sentence embeddings against which the performance of the sequence to sequence model is compared.

Since the hyperbolic vector space has different mathematical properties, a set of equations different from those used for the Euclidean space is required for the pooling methods.

$$ASE_{(i)} = \sum_{j=1}^{n} \sum_{k=1}^{300} WE_k \qquad (5)$$

$$ASE_{(i)} = \sum_{j=1}^{n} \sum_{k=1}^{300} WE_k \qquad (6)$$

$$ASE_{(i)} = \sum_{j=1}^{n} \sum_{k=1}^{300} WE_k \qquad (7)$$

3.5.2 Sequence to Sequence Model

This sentence embedding mechanism follows the sequence to sequence model introduced through the work of Sutskever et al. (2014), referred to as the seq2seq model from here onwards.

The data set is randomly shuffled and a subset consisting of 400,000 data rows is used for training the encoder, decoder units.

In the original model, the encoder accepts a set of vectors which consists of the word embedding of each word in a sentence followed by the $< EOS >$ token as input and returns a context vector as the output. In order to train the model, the decoder is fed with the context vector from the encoder, with the objective of getting the $< SOS >$ token followed by the translated sentence as the final output. For our research, the output expected from the decoder is the same sentence that has been inputted into the encoder. For a given sentence, the word embedding of each word in the sentence is inputted into the Recurrent Neural Network encoder, which has a hidden layer similar in dimensions to the word embedding. Since the expected output from the RNN decoder is also the same sentence, the context vector (output of the encoder) can be considered as the sentence embedding that we are seeking.

Different sentence embeddings are thus generated using both Euclidean and hyperbolic word embeddings as inputs to the seq2seq model. For each type of word embeddings, the RNNs inside the encoder and decoder are also modified to generate different sentence embeddings. Here, GRU (Chung et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), and simple RNN models are used. The architecture of the GRU seq2seq model has been inspired by the model introduced through the work of Cho et al. (2014).

Furthermore, two different decoder structures have been used to train the seq2seq model. A simple decoder which functions as explained above and a decoder with the attention mechanism introduced in the work of Vaswani et al. (2017) are thus utilized. Both models use a teacher forcing value of 0.5 with the objective of performing better at the prediction task (Lamb et al., 2016).

The squared L2 norm between the predicted word embedding and the actual word embedding is used as the loss function for Euclidean embeddings. Equations 8 shows embedding value of the ith sentence which is calculated by summing up all the word embeddings in the predicted sequence of word embeddings. The symbol n is the length of the longest sentence which may vary for the selected data set. WE_k is the value of each dimension in the 200 dimension word embedding. Equation 9 calculates the value of ith true word embedding sequence (TV_i) which is the summation of the word embeddings of True word sequence. Then in the equation 10, the squared L2 norm (Err) is calculated. n denotes the number of data items used. The procedure follows for both Euclidean and hyperbolic space embeddings.

$$PV_i = \sum_{j=1}^{n} \sum_{k=1}^{200} WE_K \tag{8}$$

$$TV_i = \sum_{j=1}^{n} \sum_{k=1}^{200} WE_K \tag{9}$$

$$Err = (1/n) \sum_{i=1}^{n} (PV_i - TV_i)^2 \tag{10}$$

3.6 Testing

To the extent of our knowledge, there does not exist a well known or effective benchmark to test the performance of Sinhala sentence embeddings. Therefore, the GRU RNN model with a CNN layer introduced by the work of Chung et al. (2014); Senevirathne et al. (2020) is used to test each embedding. The function of this model is to understand the sentimental reactions of Facebook users to Facebook posts and thus classify each post as either positive or negative based on its prediction of the sentimental reaction of users to that post. The classification of the Facebook posts was done as explained in the section 3.3.

As mentioned above since the scarcity of a large enough data set for Sinhala language to train deep learning models, the same Facebook data set is used for the model training purpose. However, a different set of Facebook posts are used in order to avoid repetition of the data set and a total of 200,000 posts were used for the training purpose. The holdout method was used with data set splits into the 8:1:1 ratio for train, validate, and test sets. Tests were run multiple times and the average performance measures were recorded.

4 Results

The results obtained by training the models only using word embeddings are displayed in table 2. Here, the row fastText(Sinhala News Comments) taken from the work of Weeraprameshwara et al. (2022) is used as a benchmark against which the performance measures of the other word embeddings are compared. There, the Facebook data set was embedded using the fastText word embeddings trained with the Sinhala News Comments data set introduced through the work of Senevirathne et al. (2020), while the latter rows display the results of embedding the Facebook data set with word embeddings trained with the Facebook data set itself.

As the table portrays, using the Facebook data set containing 542,871 preprocessed Facebook posts, which is much larger in size than the Sinhala News Comments data set with 15,000 Sinhala News comments, to develop the word embeddings has resulted in a comparatively higher F1 score.

The results of each embedding in the two-tiered structure are shown in table 3. The first column presents the word embedding method used while the second column depicts the sentence embedding method utilized and the rest of the columns are used to present the performance measures. The best performance measures from each word embedding category are highlighted.

The best F1 score is produced by the two-tiered embedding which uses fastText as the word embedding and the seq2seq model with GRU RNNs and attention layer as the sentence embedding while the second-best F1 is scored by the fastText embedding with average pooling. For each of the sentence embedding methods, the highest F1 score is produced by pairing with fastText word embeddings. fast-

Word Embedding	Performance Measures			
	Accuracy	Precision	Recall	F1
fastText (Sinhala News comments) (Weerapmeshwara et al., 2022)	81.17	81.17	81.57	81.37
Word2vec	83.47	83.65	83.47	83.56
GloVe	82.09	81.91	82.65	82.28
fastText	**83.76**	**83.76**	**83.76**	**83.76**
Hyperbolic	82.78	82.11	83.58	82.84

Table 2: Word Embedding results

Text embeddings have resulted in a better F1 score in the one-tiered embeddings as well. Thus, we can conclude that fastText is the word embedding schema which provides the best performance in this context.

Upon taking the word embedding categories into consideration, Word2vec embeddings provide the second-best results, with performance scores slightly lower than those of fastText. The ranking of F1 scores achieved by hyperbolic and GloVe embeddings seem to be highly dependent on the type of sentence embedding used. However, the best F1 score obtained by hyperbolic embeddings, which was by pairing with the seq2seq model with GRU encoder and decoder units including an attention layer, is higher than the best F1 score GloVe embeddings have achieved upon pairing with max pooling sentence embeddings. It should be noted that the structure of data utilized here may not be optimal for hyperbolic embeddings.

In sentence embeddings, the performance of seq2seq model with GRU encoder, decoder units and an attention layer tend to surpass other sentence embedding models except when the word embedding utilized is GloVe. Nonetheless, stripping off the attention layer brings the performance of the seq2seq model with LSTM encoders and decoders to a level higher than that obtained with GRU encoder and decoder units, with the exception of the case where hyperbolic word embeddings are utilized.

Furthermore, there is a clear improvement in the performance scores of the seq2seq model when the attention layer is applied to the decoder. However, when the attention layer is not applied, pooling embeddings manage to perform better than seq2seq models except when hyperbolic word embeddings are utilized. The reason for this exception could be that the Euclidean pooling mechanisms used may not be the best fit for hyperbolic embeddings.

5 Conclusion

Comparing tables 2 and 3 makes it evident that there is a clear improvement in performance when two-tiered embedding systems are used, in contrast to simply using a single tier of word embeddings. The possibility of sentence embeddings used in two-tiered embedding systems to enable the models to consider the syntax of sentences could be the reason for this improvement. When word embeddings of Sinhala Facebook posts are directly fed to a sentiment analysis model, the model is likely to see the Facebook posts as merely an unorganized set of words instead of an organized set of sentences.

In addition, the results displayed in table 3 exhibit the use of the two-tiered embedding system that combines fastText word embeddings and seq2seq sentence embeddings with GRU encoder and decoder units as well as an attention layer has given rise to the best performance measures. Although Word2vec embeddings follow closely behind in performance, they have failed to surpass fastText, possibly due to the inability of the embedding system to consider the internal structure of words, which the fastText embedding system by nature is capable of (Bojanowski et al., 2017; Joulin et al., 2016).

Though the hyperbolic space has an advantage over the Euclidean space due to its ability to effectively represent complex hierarchical data structures (Nickel and Kiela, 2017), fastText and Word2vec have outperformed hyperbolic embed-

Embedding level		Performance Measures			
Word	**Sentence**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Word2vec	Max Pooling	77.23	80.06	94.59	86.72
	Min Pooling	77.29	**81.55**	92.41	86.64
	Avg Pooling	77.44	81.43	93.41	87.01
	Seq2seq GRU	75.86	76.74	97.14	85.75
	Seq2seq GRU with attention	**79.12**	79.72	96.45	**87.29**
	Seq2seq LSTM	75.97	76.17	**98.76**	86.01
	Seq2seq LSTM with attention	77.42	77.86	97.36	86.53
GloVe	Max Pooling	75.63	77.74	96.79	**86.22**
	Min Pooling	75.34	**78.68**	94.81	85.99
	Avg Pooling	**76.11**	76.90	97.38	85.93
	Seq2seq GRU	74.23	74.15	**100.00**	85.16
	Seq2seq GRU with attention	74.23	74.09	**100.00**	85.12
	Seq2seq LSTM	74.23	74.15	**100.00**	85.16
	Seq2seq LSTM with attention	74.23	74.09	**100.00**	85.12
fastText	Max Pooling	79.93	81.23	94.78	87.49
	Min Pooling	79.80	82.49	93.22	87.52
	Avg Pooling	**80.86**	**82.55**	94.07	87.93
	Seq2seq GRU	78.12	80.90	92.33	86.23
	Seq2seq GRU with attention	80.61	81.31	96.00	**88.04**
	Seq2seq LSTM	79.00	82.12	91.59	86.60
	Seq2seq LSTM with attention	80.31	80.06	**96.98**	87.72
Hyperbolic	Max Pooling	76.71	77.54	95.95	85.77
	Min Pooling	76.11	77.68	94.11	85.11
	Avg Pooling	77.00	77.31	95.56	85.47
	Seq2seq GRU	76.38	77.09	**97.57**	86.13
	Seq2seq GRU with attention	**77.31**	**78.31**	96.70	**86.54**
	Seq2seq LSTM	76.48	77.91	95.49	85.81
	Seq2seq LSTM with attention	77.19	78.22	96.24	86.30

Table 3: Performance measures of each embedding

dings in this research. The reason for this could be the lack of potent parsing tools for the Sinhala language (de Silva, 2019). To obtain the optimum performance from hyperbolic embeddings, an effective hierarchical structure such as sentence structures identified via parsing is required. The simple

$[word, sentence]$ relation structure used in this research may not be sufficient for this. Furthermore, the pooling techniques also fail to be on par with the seq2seq model, possibly due to the fact that the vectors generated by applying Euclidean pooling mechanisms on hyperbolic embeddings do not always fall within the space of the Poincaré ball.

Another noteworthy fact is that the GloVe embeddings tend to underperform in comparison to the other word embeddings models used in this research. Unlike resource-rich languages such as English, no pre-trained GloVe models exist for the Sinhala language. This could hinder the ability of GloVe embeddings to achieve their full potential.

Thus, it can be concluded that though a robust embedding model for Sinhala that is applicable across all domains may not be currently available, it could be possible to develop an effective embedding system that would at least be potent within the domain of the training data set by applying a two-tiered embedding model such as the seq2seq sentence embeddings with GRU encoders and decoders stacked on top of fastText word embeddings on a sufficiently large data set.

6 Future Work

This research is related to the work of Jayawickrama et al. (2021) and as the final goal, a Facebook reaction prediction tool for colloquial Sinhala text will be developed and the word representations developed in this project will be used for that tool.

The data set contains both Sinhala and English text since our aim is to develop a word representation for colloquial Sinhala text which consists of English and Sinhala code-mixed content. However, a pure Sinhala embedding can be generated in the future.

Furthermore, Poincaré embeddings could be developed for Sinhala text with the use of a proper parser to identify sentence structures though developing a reasonable parser for colloquial text will be a challenge.

Though this research only considers sentiment analysis for the Sinhala language, the applicability of the two-tiered embedding systems discussed in other areas of natural language processing as well as for other resource-poor languages could be tested as well.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. 2017. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Nisansa de Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. 2018. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*.

Doaa Mohey El-Din. 2016. Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 7(1).

Cole Freeman, Hamed Alhoori, and Murtuza Shahzad. 2020. Measuring the diversity of facebook reactions to research. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–17.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

Lisa Graziani, Stefano Melacci, and Marco Gori. 2019. Jointly learning to detect emotions and predict facebook reactions. In *International Conference on Artificial Neural Networks*, pages 185–197. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Vihanga Jayawickrama, Gihan Weeraprameshwara, Nisansa de Silva, and Yudhanjaya Wijeratne. 2021. Seeking sinhala sentiment: Predicting facebook reactions of sinhala posts. In *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, pages 177–182. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

AB Kanduboda. 2011. The role of animacy in determining noun phrase cases in the sinhalese and japanese languages. *Science of words*, 24:5–20.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in neural information processing systems*, pages 4601–4609.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Matthias Leimeister and Benjamin J Wilson. 2018. Skip-gram word embeddings in hyperbolic space. *arXiv preprint arXiv:1809.01498*.

Qiuhao Lu, Nisansa de Silva, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Berthold Reinwald, and Yunyao Li. 2020. Exploiting node content for multiview graph convolutional network and adversarial regularization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 545–555.

Qiuhao Lu, Nisansa de Silva, Sabin Kafle, Jiazhen Cao, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Brent Hailpern, Berthold Reinwald, and Yunyao Li. 2019. Learning electronic health records through hyperbolic embedding of medical ontologies. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 338–346.

Taylor Lyles. 2020. Facebook adds a 'care' reaction to the like button.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2016. Geometry of polysemy. *arXiv preprint arXiv:1610.07569*.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30:6338–6347.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. *arXiv preprint arXiv:1611.02988*.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR.

Robert E Schapire and Yoav Freund. 2012. Foundations of machine learning. *Mit Press*.

Lahiru Senevirathne, Piyumal Demotte, Binod Karunanayake, Udyogi Munasinghe, and Surangika Ranathunga. 2020. Sentiment analysis for sinhala language using deep learning techniques. *arXiv preprint arXiv:2011.07280*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2428–2437.

Gihan Weeraprameshwara, Vihanga Jayawickrama, Nisansa de Silva, and Yudhanjaya Wijeratne. 2022. Sentiment analysis with deep learning models: a comparative study on a decade of sinhala language facebook data. In *2022 The 3rd International Conference on Artificial Intelligence in Electronics Engineering*, pages 16–22.

Yudhanjaya Wijeratne and Nisansa de Silva. 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. *arXiv preprint arXiv:2007.07884*.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.

Junhao Zhou, Yue Lu, Hong-Ning Dai, Hao Wang, and Hong Xiao. 2019. Sentiment analysis of chinese microblog based on stacked bidirectional lstm. *IEEE Access*, 7:38856–38866.

V-collocates with *will* and *be going to*:
A Corpus-based Analysis

Jarry Chia-Wei Chuang
Department of English
National Chengchi University
Taipei, Taiwan
cwchuang.academia@gmail.com
ORCID: 0000-0002-3029-4463

Abstract

A corpus-based study is conducted to investigate verbs collocating with present-tensed *will* and *be going to*. Data from Corpus of Contemporary American English (COCA) are addressed, especially for lexical verb-collocates (V-collocates). Data-driven results show that lexical V-collocates of *will* and *be going to* are sensitive to semantic and pragmatic functions. MI scores manifest that verbs with deficient meanings but higher functionality can be commonly applied to *will* contexts. V-collocates in *will* contexts have a wider acceptance of register, despite a potential stylistic preference to formal use; in contrast, V-collocates in *be going to* reveal its lower formality but higher frequency in informal spoken context. Additionally, semantic expansion with metaphorical and hilarious use is interestingly found in informal *be going to* contexts, especially on words involved with serious legal or security issues. In general, meaning shifts within *will* and *be going to* are greatly influenced by and pragmatically derived from context, which is coordinated with the monosemous account by Nicolle (1998a). The present findings offer new information about the distinction of English future event markers.

Keywords: verb collocates (V-collocates), will, be going to, corpus-based, COCA

1 Introduction

1.1 Overview

As primary English future event expressions, *will* and *be going to* have been long explored for their intricacies in syntactic, semantic, and pragmatic aspects (Berglund, 1997, 2005; Haegeman, 1989; Nicolle, 1998a; Szmrecsanyi, 2003). Researchers have demonstrated their efforts in the distinction with different methods and from theoretical aspects, while the relation between *will* and *be going to* in actual use remains complicated.

1.2 Research Difficulties

The difficulties in distinguishing them can be derived from (1) their intimate associations with tense, aspect, and mood/modality (TAM), (2) their sensitivity to syntactic and semantic conditions, and (3) disambiguation from their semantics or pragmatics. As Haegeman (1989) noted: “to analyze future markers in English, we address not only tense and aspectual auxiliaries but also the relation between surface representations of tense and truth conditions.” In addition, by examining syntactic environments, Szmrecsanyi (2003) further indicates that English future event markers are sensitive to negation, subordinate contexts, IF-clause, and even sentence length. As for their disambiguation, a monosemous account was proposed by Nicolle (1998a), in which the varieties of *will* and *be going to* interpretations and their function to express future events are pragmatically derived. Multiple factors have been widely discussed, suggesting that follow-up studies should better consider those intervenient elements at different linguistic levels.

1.3 Research Gap

Previous discussions mainly focused on the complementizer layer and inflectional layer (CP & IP), while attention was seldom paid to the lexical layer (vP & VP). As the meaning core of subject predicates, sentential expressions, and utterances, lexical verbs (V_{LEX}) (with lower functionality but rich semantic knowledge), in comparison with *will* and *be going to*, should provide vital information. Verbs collocating with *will* and *be going to* are thus believed to play a significant role in explaining their meanings as well as marking the differences in use.

1.4 Purpose of Study

The present study investigates present-tensed *will* and *be going to*, with an innovative point of view. We explore the pattern at the lexical layers, by scrutinizing their verb collocates (V-collocates). Research Data is retrieved from the Corpus of Contemporary American English (COCA). V-collocate differences are expected to show different semantic and pragmatic functions of the two markers. It is worth noting that considering the sensitivity to TAM as well as the complex interaction with tense and aspectual auxiliaries. So, in the paper, we mainly discuss lexical V-collocates (*will* / *be going to* + V_{LEX}).

2 Syntax

Some of the previous studies on *will* and *be going to* attach importance to their syntactic environments. To observe their similarity and difference in syntactic patterns, they may show implications on the distinction. Those attempts done previously can be summarized and classified in the 3-layered structure of syntax.

Rizzi (1997) categorizes the syntactic structure into three layers, including the complementizer layer (CP), inflectional layer (IP), and lexical layer (vP & VP). Associating syntactic layers with previous discussions, we find that studies on the CP and IP levels were mostly explored previously, while the lexical layer in *will* and *be going to* contexts was seldom under inspection.

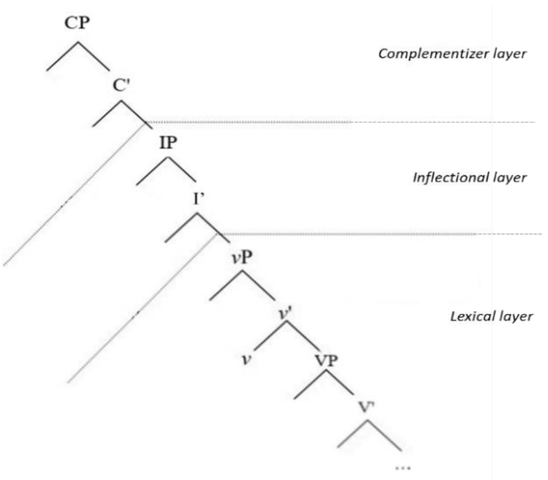


Figure 1. The three-layered structure of English syntax (Rizzi, 1997)

2.1 Complementizer layer (CP)

The complementizer layer is generally headed by a functional word or morpheme (e.g., *for*, *since*, *when*, *if*, *as*, etc.). Type of sentence is also determined here. CP-level discussions about *will*-and-*be-going-to* distinction include their sensitivity to subordinate contexts and IF-clause.

In Szmrecsanyi's (2003) study, *be going to* presents a high-frequency pattern in syntactic dependent contexts (i.e., subordinate contexts) but a relatively lower frequency in independent contexts (i.e., the main clauses of sentences); on the other hand, reservations are held for *will* contexts, since *will* seems not to demonstrate an obvious tendency towards syntactic independence within sentences. In general, subordinate contexts have perceptible influences on the occurrence of *be going to*. Moreover, Yeh (2021) reexamined subordinate structures in both contexts, using data from the British National Corpus (BNC). It is indicated that in the BNC data, *will* is much more frequent in independent contexts. Note that Szmrecsanyi's (2003) data sources embraced British English (BE) and American English (AE), while Yeh (2021) only adopted a BE-based corpus. This infers the use of *will* and *be going to* may have dialectal variations and cross-cultural bias, in which AE and BE speakers show different degrees of preference.

Szmrecsanyi (2003) also pointed out that *will* and *be going to* have extraordinary sensitivity in if-subclause (IF-Sub) structure, among subordinate constructions: *will* is more dominant in main clauses of IF-Sub clauses; *be going to* is frequent in IF-Sub clauses. IF-Subs usually lead conditional sentences, where their truth values are opaque or unsure. Conditional sentences often coordinate with mood and modality to avoid over-affirmation. Divergent distributions in IF-Subs confirm that *will* and *be going to* are tied to TAM.

2.2 Inflectional layer (IP)

The inflectional layer is often headed by functional heads. According to the split-IP hypothesis (Pollock, 1989), IP can be divided into several sub-layers, some of which may be subtly different across languages. Verb can move from lower phrases to TP, which is called verb raising. In English syntax, one of the sub-phrases in the split IP is Negation Phrase (NegP).

In previous studies, negation is reported to be influentially associated with the distribution of *will* and *be going to* (Szmrecsanyi, 2003) in the IP level: *won't* frequently occur in negated contexts, while data excluding *won't* illustrates that *will* is hardly negated. As for *be going to*, data of its negation does not reveal an explicit preference. This phenomenon should be noted by the analogical pressures in AE, in which AE speakers prefer contracted forms (cf. Hofland & Johansson, 1982; Hundt, 1997; Szmrecsanyi, 2003); thus, the frequency of contracted *won't* in AE can be higher than that of *not going to*. To study *will* and *be going to*, cross-dialectal difference should be particularly noted.

2.3 Lexical layer (vP & VP)

The lexical layer is normally headed by the verb, in which the argument structure and theta roles work. In the syntax of *will* and *be going to*, the main IPs are occupied by *will* and *be going*. V_{LEX} serves as the head of vP, as demonstrated in (1). The syntax of *will* is less complicated than *be going to*, in that V_{LEX} in the latter context is located at the vP of another lower-layered CP.

(1) Syntax of present-tensed *will* and *be going to*

a. *will* + V_{LEX}:
$[_{TP}$ *will* $[_{vp}$ V_{LEX} …

b. *be going to* + V_{LEX} :
$[_{TP}$ *is/am/are* $[_{AgrP}$ *going* $[_{CP}$ $[_{TP}$ *to* $[_{vP}$ V_{LEX} …

Literature about *will* and *be going to* mostly discuss their behaviors at CP and IP layers, hardly coping with the syntax at the lexical layer (vP & VP); however, V_{LEX} following them can provide vital information in the distinction, since V_{LEX} can not only be C-selected but also be S-selected by *will* and *be going to*.

It would be argued here that V_{LEX} is worth noting and might be more truthful to capture the pattern of *will* and *be going to*, since it is dominated at the lower layer.

3 Semantic-Pragmatic Interactions

Aside from the syntactic distributions, the semantics and pragmatics of *will* and *be going to* should also be examined and well defined first. They seem to be polysemous in surface, in which *will* is more than expressing volition and *be going to* is more than signifying movement. Checking the evolution of their meanings, we might find a successive process of meaning expansion and they might not be polysemous despite multiple meanings that have been derived in both contexts.

As the semantic model of English modal auxiliaries proposed by Klinge (1993), Nicolle (1997) argued that *will* and *be going to* should be analyzed as monosemous, under the Relevance Theory. Even though users usually think of more than one possibility that *will* and *be going to* can derive, this does not make their polysemy. Instead, such expressive diversity of *will* and *be going to* is considered to be pragmatically derived (Nicolle, 1998a). The apparent polysemy is caused by their context-sensitive interpretations, highly associated with grammaticalization (i.e., shifting from heavy verbs to light verbs/modal auxiliaries) (Nicolle, 1998b). Grammaticalized *will* and *be going to* make themselves seemingly polysemous. This can be attributed to semantic retention. Original senses are kept but only triggered in certain contexts, in which lexical meanings and grammaticalized senses co-exist. So, *will* and *be going to* offer several potential interpretations in contexts and those should be interpreted as sense derivation at pragmatic levels.

3.1 *will*

Commonly known as modal auxiliaries in English, *will* derives several contextual meanings, appearing to state (1) possibility with present or future time references, (2) habitual expressions, (3) willingness to make performance or take action, and so on. Previous analyses have drawn the monosemous analysis and attributed such polysemous interpretations to semantic retention; a single sense is actually taken. Before grammaticalized, *will* represents its old (lexical) meaning to express one's volition. But, inferential meanings become accessible from the relevance-theoretic perspective. Interpretations through three major functions of English modal auxiliaries (e.g., epistemic, deontic, and dynamic) hence expand its semantically underdefined sense, leading to various contextual meanings.

3.2 *be going to*

In use, *be going to* reflects activities in the upcoming future or inevitability (Coates, 1983; Leech, 2014; Palmer, 2014). Two different meanings are usually considered: (1) prior intention, and (2) current activities (Nicolle, 1998a). Distinct default senses challenge the monosemous account for *be going to,* while it is solved by semantic retention as well. The default meaning is found to be kept and proved evident in native children processing their L1 (English) (Ziegeler, 1996, 1997). Polysemous interpretations are in fact by-products of grammaticalization. V-collocates thereby have the possibility to trigger such contextually polysemous meanings in *be going to* conditions. Brisard (2001) further notes that *be going to* can draw a "paradoxical but pragmatically plausible" account for the near future, in which events remain unknown in the present timing usually, before declared by the speaker.

3.3 Importance of the Monosemy

In general, the semantics of *will* and *be going to* can be treated as monosemy, though. Monosemy should be noted before the following analysis of corpus data. As *will* and *be going to* are monosemous and are semantically based on a single sense, V-collocates or the whole VP can take the floor to greatly influence and even determine the sentence and contextual meanings. Then, V-collocates gain the importance to be checked and discussed in the following sections.

4 Method

The present analysis relies on corpus data from native users' examples, which is believed to better connect close-to-nature patterns to the distinction between *will* and *be going to.*

4.1 Data Source

Research data under close scrutiny are consulted from the Corpus of Contemporary American English (COCA) (Davies, 2008), one of the most representative corpora of American English (AE). It comprises over 1 billion words and a wide array of sentences in genres with various formality, from spoken text to academic context. COCA also provides instant operation of frequency and MI value.

4.2 Procedure & Data Analysis

To explore V-collocates with *will* and *be going to*, we will first categorize types of V-collocates, since they can potentially be attached by lexical verbs or tense-aspectual auxiliaries. Modified data have been under operation and checked by two well-trained research fellows. Two V-collocates with English future markers are examined through frequency ($\mathrm{Min}_{\mathrm{freq}} = 20$) and MI scores ($\mathrm{Min}_{\mathrm{MI}} = 3$).

5 Results & Discussion

V-collocates may include lexical verbs and auxiliary verbs (e.g., progressive auxiliary, passive auxiliary, perfect auxiliary, etc.). Raw data consulted from COCA should be under re-analysis to scout out lexical V-collocates. V-collocates can further be explored via frequency and MI scores.

5.1 Type of V-Collocates

Verbs attached to *will* and *be going to* are miscellaneous. Note that BE-verbs should be cautiously examined, for they can serve as (1) intensive verbs proceeded by subject predicates like adjectives or nouns, (2) progressive auxiliaries [PROG] followed by Ving, and (3) passive auxiliaries [PASS] to express passive voice. Besides, *have* also needs close examination, since it can be a lexical verb as well as the perfect auxiliary [PREF]. So, the data presented below has been amended based on raw frequency excerpted from COCA. Four types of V-collocates dominating most of the uses are shown as follows:

	will	*‘ll*	**Total**	**Occurrence**
X + V$_{LEX}$	1,480,274	835,879	2,316,153	90.81%
X + be Ving	39,687	29,190	68,877	2.70%
X + be Vpp	140,797	15,515	156,312	6.13%
X + have Vpp	8,200	1,105	9,305	0.36%
Total	1,480,274	835,879	2,316,153	100.00%

Note: X represents two different forms of WILL, *will* or ‘ll.

Table 1: Type of V-collocates with *will*

	is	*‘s*	*am*	*‘m*	*are*	*‘re*	**Total**	**Occurrence**
Y + V$_{LEX}$	78,467	91,977	7,403	65,106	56,010	126,957	182,967	88.63%
Y + be Ving	1,823	1,689	174	1,192	2,512	4,431	11,821	5.73%
Y + be Vpp	4,046	2,346	57	462	2,917	1,745	11,573	5.61%
Y + have Vpp	14	14	-	6	19	34	87	0.04%
Total	84,350	96,026	7,634	66,766	61,458	133,167	206,448	100.00%

Note: Y stands for different forms of BE going to, in which BE can be *is* (*‘s*), *am* (*‘m*), or *are* (*‘re*).

Table 2: Type of V-collocates with *be going to*

In terms of frequency, corpus data illustrates *will* is more prevailing than *be going to*. V-collocates of present-tensed *will* and *be going to* in **Table 1&2** reveal that they are regularly attached to lexical verbs ($N_{1,\ LEX}$=2,316,153; $N_{2,\ LEX}$=182,967). In addition, BE-verbs collocating with two verbs demonstrate subtly different patterns, especially when they are progressive auxiliaries and passive auxiliaries. Collective frequencies of passive and progressive use in *will* context display the frequency of passive use (be+Vpp) ($N_{1,\ be+Vpp}$=140,797) is divergently higher than progressive use (be+Ving) ($N_{1,\ be+Ving}$=39,687) in *will*, while progressive use (N_1=29,190) overwhelms passive use (n=15,515) in cliticized *’ll* conditions. Progressive-passive biased use is not obviously found in *be going to* context ($N_{2,\ PROG}$=11,821; $N_{2,\ PASS}$=11,573).

5.2 Frequency of V-collocates

Results of V-collocates by frequency manifest a wide variety of actions that seem not to be semantically associated with *will* and *be going to*. In **Table 3**, high-frequency verbs collocating with present-tensed *will* and *be going to* are displayed. Note that in present-tensed *will* and *be going to* contexts, frequency of *be* is counted, as it serves as an intensive verb to bring the subject predicates out. Yet, V-collocates presented as be+Ving and be+Vpp have been excluded.

will$_{[PRES]}$	*be, have, take, make, get, do, continue, come, help, go, see, find, give, need, become, tell, say, happen, work, keep*
be$_{[PRES]}$ going to	*be, have, get, do, take, go, make, happen, see, come, need, give, say, try, tell, talk, start, put, find, look*

Table 3: V-collocates with present-tensed *will* and *be going to* by frequency

The patterns of V-collocates are analogous to each other. First, *be*, as a lexical verb, is attached to both most frequently. Secondly, verbs carrying abundant meanings (e.g., *have, get, make, take, do, go*, etc.) are frequently attached to both future markers. This could be taken as part of the evidence that *will* and *be going to* share similar functions in identifying the high probability of the occurrence.

5.3 MI Scores & V-collocates

As frequency patterns between *will* and *be going to* provide limited information on the distinguishment, the following section will analyze MI scores of V-collocates. Four dominant genres are listed in **Table 4&5**, in which the majority of the register distribution can be found: (1) spoken, (2) magazine, (3) newspaper, and (4) academic contexts. Formality will also be counted based on the sum of V-collocates in most of the formal contexts (i.e., magazine, newspaper, and academic contexts). To avoid selection bias, the ratios of occurrence in all the selected contexts are calculated, with the time of use in all documented contexts as the denominator. Results of mutual information reveal a variety of differences in V-collocates. It is found that V-collocates have intimate interaction with semantic and pragmatic functions, including meanings of lexical verbs, genres, and formality.

5.3.1 *will*

V-collocate	MI Score	All	Informal		Formal						Formality (M+N+A)/All	
			Spoken		Magazine		Newspaper		Academic			
		N	N	%	N	%	N	%	N	%	N	%
continue	7.42	25,413	4,348	17.1%	3,075	12.1%	4,306	16.9%	2,832	11.1%	10,213	40%
prevail	7.14	827	214	25.9%	98	11.9%	97	11.7%	53	6.4%	248	30%
enable	7.05	2,152	115	5.3%	509	23.7%	234	10.9%	586	27.2%	1,329	62%
depend	6.99	2,493	209	8.4%	425	17.0%	379	15.2%	642	25.8%	1,446	58%
remain	6.70	8,038	779	9.7%	1,160	14.4%	1,635	20.3%	1,059	13.2%	3,854	48%
require	6.48	5,799	271	4.7%	1,042	18.0%	839	14.5%	1,624	28.0%	3,505	60%
receive	6.29	5,240	255	4.9%	582	11.1%	1,309	25.0%	536	10.2%	2,427	46%
allow	6.27	8,386	809	9.6%	1,390	16.6%	1,323	15.8%	1,144	13.6%	3,857	46%
emerge	6.14	1,075	90	8.4%	243	22.6%	191	17.8%	197	18.3%	631	59%
happen	6.04	11,608	2,359	20.3%	1,113	9.6%	1,295	11.2%	425	3.7%	2,833	24%
affect	5.97	3,007	397	13.2%	456	15.2%	494	16.4%	511	17.0%	1,461	49%
begin	5.92	6,892	882	12.8%	948	13.8%	1,829	26.5%	469	6.8%	3,246	47%

Table 4: Lexical verbs collocating with present-tensed *will* by MI score.

V-collocates with *will* have an inclination toward verbs like (1) intensive verbs, (2) prepositional (or intransitive) verbs, and (3) verbs proceeded by non-finite clauses. Such verbs require subject-predicates or object-predicates to make the VPs semantically complete. For example, *remain* serves as an intensive verb and requires a subject-predicate, such as NP or AP, to complete the meaning of the whole VP. Besides, intransitive verbs or prepositional verbs also usually attach to *will*, including *begin*, *continue*, *depend*, *emerge*, *happen*, etc. They can either exist independently in intransitive use or look for an infinitive to complete the meaning of the whole VP. What's more, verbs like *allow*, *enable*, and *require* are commonly followed by a non-finite clause to complement the expressions. See also *affect*, as in (2a) and (2b). *Affect* is a semantically fuzzy word like *influence* or *impact*. It infers that the prediction is uncertain to some extent. If there is a high probability to occur, speakers should know more details and would rather take more informative verbs to elaborate on how the situation can be influenced. In sum, V-collocates in *will* contexts seem to be semantically defective and possess higher functionality.

(2a) Your vote will ***affect*** the future and be recorded in eternity. (BLOG, 2012)
(2b) […], which I also don't think will ***affect*** many people. (NEWS, 1990)

Pragmatically speaking, V-collocates with *will* appears more often in formal contexts. Checking the tendency of formality, we find most of the verbs present a relative increase in frequency, transitioning from spoken context to academic contexts. For example, *enable* and *depend* only show 5.3% and 8.4% occurrence in spoken contexts, but their occurrences obviously surge to 27.2% and 25.8% in academic contexts. Similar patterns can be found in other V-collocates with *will*. Most of the cumulative occurrence rates (including use in magazines, newspapers, and academic contexts) reach 40% and more. Yet, the formality distributions mostly hover around 40-60%. It reveals *will* has a wider acceptance of different registers to some extent (cf. formality value of *be going to* in **Table 5**), in spite of the potential preference towards formal contexts.

5.3.2 *be going to*

V-collocate	MI Score	All	Informal		Formal						Formality (M+N+A)/All	
			Spoken		Magazine		Newspaper		Academic			
		N	N	%	N	%	N	%	N	%	N	%
happen	7.01	17,571	7,568	43.1%	989	5.6%	1,621	9.2%	186	1.1%	2,796	16%
die	6.06	5,619	1,230	21.9%	429	7.6%	359	6.4%	73	1.3%	861	15%
explode	5.95	335	58	17.3%	23	6.9%	13	3.9%	2	0.6%	38	11%
kill	5.55	4,756	1,034	21.7%	194	4.1%	243	5.1%	39	0.8%	476	10%
jail	5.50	1,403	328	23.4%	60	4.3%	104	7.4%	16	1.1%	180	13%
lose	5.38	3,347	1,320	39.4%	220	6.6%	337	10.1%	27	0.8%	584	17%
win	5.35	4,885	2,095	42.9%	318	6.5%	617	12.6%	24	0.5%	959	20%
marry	5.34	841	164	19.5%	52	6.2%	38	4.5%	2	0.2%	92	11%
be	5.26	156,090	70,649	45.3%	8,903	5.7%	15,274	9.8%	1,751	1.1%	25,928	17%
retire	5.20	315	135	42.9%	22	7.0%	63	20.0%	1	0.3%	86	27%
solve	5.20	897	445	49.6%	62	6.9%	99	11.0%	13	1.4%	174	19%
continue	5.04	3,702	2,310	62.4%	151	4.1%	366	9.9%	46	1.2%	563	15%

Table 5: Lexical verbs collocating with present-tensed *be going to* by MI score.

As shown in Table 5, *die*, *explode*, *kill*, and *jail* are verbs involved with serious legal or security issues. Normally, speakers would not address such verbs for sensitivity, especially when things have been known to certainly happen. Referring to the original sense of *be going to*, we find its V-collocates in part related to the meanings: (1) events in the near future and (2) inevitability. The distributions are attributed to the latter one. While it is confirmed to have occurrence and display the certainty, those verbs can be linked to the use of *be going to* (i.e., pure future and inevitability) much better than *will*. The rules can likewise explain other V-collocates like *lose*, *win*, *marry*, and *retire*. Those verbs are used when the happening of the action or the appearance of status is for sure for speakers.

Furthermore, an intriguing use with semantic expansion is frequently found in *be going to* context. For example, *explode* and *kill* are commonly associated with legal or security issues. As they are thrown into the spoken context, the original meaning of *explode* and *kill* are weakened; yet, metaphorical use and humorous expressions take place to expand their uses. Look at the examples below.

(3a) Bob is going to ***explode***. (SPOK, 2014)
(3b) I believe you, Joe. […] You're not going to ***kill*** me? (SPOK, 2017)

In (3a), it would be generally reckoned that Bob is going to be angry, instead of breaking up into pieces. The speaker describes that Bob's emotions would erupt angrily, dangerous as a bomb. As for

(3b), where a question is hilariously asked. By observing the context, it can be easy to understand that the speaker did not really confirm whether his friend was going to kill him, but made fun of his friends and try to tell his friend not to be angry. The speaker covertly manipulates the cooperative principle, flouting the maxims. In fact, using *kill* to play the boundary is a typical example, in that many examples do not derive *kill* from “cause to die” (cf., Fodor, 1970; Wierzbicka, 1975).

Though semantic expansion seemingly takes place, such processes are pragmatically derived. In (3a) and (3b), word senses seem to be humorously or metaphorically applied to informal contexts. However, it should be noted such meanings originally occur in specific contexts. Expansion of those word senses takes place, as the meanings have been conventionalized under valid inference for a long time. So, (3a) and (3b) should be known as the transition for their pragmatics to semantics.

With respect to genres and formality, V-collocates with *be going to* have a preference for informal registers. Generally, they share a high frequency in spoken contexts rather than formal registers. Most of the formality distributions are lower than 20%, which implies that *be going to* may be commonly used in informal contexts or daily use, in comparison to *will*.

5.3.3 Comparison between *will* & *be going to*: *continue* & *happen*

Comparing *will* and *be going to*, we further found *continue* and *happen* cooccur in the top10 V-collocates lists by MI score, with high MI scores in both. For their high accessibility, this section is proposed to comparatively discuss their behaviors in both contexts. They are especially noticeable in frequency as well as MI score (see **Table 6**).

VLEX	*will*	*be going to*
continue	7.42 (25,413)	5.04 (3,702)
happen	6.04 (11,608)	7.01 (17,571)

Note: Numbers in the table refer to MI Score (Frequency).

Table 6: Representation of *continue* and *happen* by MI score and frequency

Considering their MI scores and frequencies, we found *continue* is preferred in *will* contexts than *be going to* contexts, with higher MI scores and frequencies (MI: 7.42 > 5.04; Freq: 25,413 > 3,702). In contrast, *happen* has been used more in *be going to* contexts (MI: 6.04 < 7.01; Freq: 11,608 < 17,571). The numbers of *happen* is less divergent, which needs further inspection. Following the previous section, we summarized their distribution of formality to check the preferences (**Table 7**).

VLEX	***will***		***be going to***	
	Informal (SPOK)	Formality	Informal (SPOK)	Formality
continue	4,348	40%	2,310	15%
	17.1%		62.4%	
happen	2,359	24%	7,568	16%
	20.3%		43.1%	

Table 7: Formality preference of *continue* and *happen* in *will* and *be going to* contexts.

V-collocates of *continue* and *happen* confirm that *will* is less preferred in informal contexts, while *be going to* bears lower formality and is used more frequently in informal contexts. Generally speaking, V-collocates with *will* and *be going to* can be sensitive to formality.

6 Conclusion

The study employs a corpus-based approach to scrutinize V-collocates with *will* and *be going to* in the present tense. Results show their V-collocates are highly associated with their semantics and pragmatics. First, MI scores manifest V-collocates with *will* are verbs with deficient meanings but higher functionality like (1) intensive verbs, (2) prepositional (or intransitive) verbs, or (3) verbs proceeded by non-finite clauses. Secondly, their V-collocates are sensitive to genres and formality. V-collocates of *will* has a wider acceptance of register than those of *be going to*; on the other hand, V-collocates with *be going to* display lower formality and higher frequencies in informal spoken context. Third, associating verb behaviors and contexts of use, we find that semantic expansion with metaphorical or hilarious use is frequently found in informal *be going to* contexts, in that figurative and analogical meanings are potentially derived from their denotations and pragmatically manipulated to be semantic expansion. This is coordinated with the monosemous account drawn by Nicolle (1998a). By and large, the present findings reveal that V-collocates with present-tensed *will* and *be going to* are sensitive to their distinction.

Acknowledgments

The study has been supported by a grant from College of Foreign Languages & Literature, National Chengchi University. Besides, the author would like to thank Prof. Huei-ling Lai and the anonymous reviewers of PACLIC-36 for their comments.

References

Berglund, Y. (1997). Future in present-day English: Corpus-based evidence on the rivalry of expressions. *ICAME journal, 21*, 7-20.

Berglund, Y. (2005). *Expressions of future in present-day English: A corpus-based approach.* Acta Universitatis Upsaliensis,

Brisard, F. (2001). Be going to: An exercise in grounding. *Journal of Linguistics, 37*(2), 251-285.

Coates, J. (1983). *The semantics of the modal auxiliaries*: Routledge.

Davies, M. (2008). The corpus of contemporary American English (COCA). In.

Fodor, J. A. (1970). Three reasons for not deriving" kill" from" cause to die". *Linguistic Inquiry, 1*(4), 429-438.

Haegeman, L. (1989). Be going to and will: a pragmatic account. *Journal of Linguistics, 25*(2), 291-317.

Hofland, K., & Johansson, S. (1982). *Word frequencies in british and american english*: Norwegian computing centre for the Humanities.

Hundt, M. (1997). Has British English been catching up with American English over the past thirty years. *Corpus-based studies in English. Amsterdam: Rodopi*.

Klinge, A. (1993). The English modal auxiliaries: from lexical semantics to utterance interpretation1. *Journal of Linguistics, 29*(2), 315-357.

Leech, G. N. (2014). *Meaning and the English verb*: Routledge.

Nicolle, S. (1997). A relevance-theoretic account of be going to. *Journal of Linguistics, 33*(2), 355-377.

Nicolle, S. (1998a). Be going to and will: a monosemous account. *English Language & Linguistics, 2*(2), 223-244.

Nicolle, S. (1998b). A relevance theory perspective on grammaticalization.

Palmer, F. R. (2014). *Modality and the English modals*: Routledge.

Pollock, J.-Y. (1989). Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry, 20*(3), 365-424.

Rizzi, L. (1997). The fine structure of the left periphery. In *Elements of grammar* (pp. 281-337): Springer.

Szmrecsanyi, B. (2003). Be going to versus will/shall: Does syntax matter? *Journal of English Linguistics, 31*(4), 295-323.

Wierzbicka, A. (1975). Why" kill" does not mean" cause to die": the semantics of action sentences. *Foundations of language, 13*(4), 491-528.

Yeh, T.-F. (2021). A corpus-based investigation of semantic and syntactic differences between the two major future tense constructions. *Concentric, 47*(1), 34-60.

Ziegeler, D. (1996). A synchronic perspective on the grammaticalisation of WILL in hypothetical predicates. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 20*(2), 411-442.

Ziegeler, D. (1997). Retention in ontogenetic and diachronic grammaticalization.

Empirical Evaluation of Language Agnostic Filtering of Parallel Data for Low Resource Languages

Praveen Dakwale[1] **Talaat Khalil**[2] **Brandon Denis**[3]
Huawei Technologies R&D
Amsterdam, Netherlands
praveen.dakwale@huawei.com[1],
khalil.talaat@gmail.com[2],
brandon.james.denis@huawei.com[3]

Abstract

Most of the available resources for low resource languages are crawled from the web. In order to obtain reasonable machine translation performance with such datasets, it is important to filter low quality samples from the training data. In this paper we explore the use of language agnostic sentence representations for filtering parallel data for low resource language pairs: Pashto-English, Khmer-English, Nepali-English and Sinhalese-English. We determine the quality of the samples based on embedding similarity between source and target sentences. Our experiments show that when preceded by language filtering using language agnostic embeddings significantly improves the performance of neural machine translation (NMT) and achieve performance competitive to language specific approaches.

1 Introduction

Neural machine Translation models are known to be data hungry (Koehn and Knowles, 2017). Training a high-quality NMT model requires a very large amount of data, usually in the order of millions of sentence pairs. For the majority of language pairs, parallel training data is compiled by aligning web-crawls in source and target languages using various heuristics based methods (Munteanu and Marcu, 2005). These web crawled datasets are often noisy due to alignment errors between source and target sentences. These misalignments lead to models with poor translation performance (Khayrallah and Koehn, 2018). Therefore, it is important to ensure the quality of the training data by filtering out noisy samples.

Various filtering techniques have been proposed in the machine translation literature which focus on different types of noise. Significant gains can be achieved by applying simple rules that do not take into consideration the semantic similarity between source and target sentences. Such rules include: removal of sentence pairs that are identified with language codes that are not aligned with source and target languages, length based pruning, removal of repetitive strings, and other heuristics (Barbu and Barbu Mititelu, 2018). However, a major type of noise is the non-equivalence of source and target sentences. Here, non-equivalence implies that the source and target sentences are not correct translations of each other (Khayrallah and Koehn, 2018). These non-equivalence cases are much harder to identify by simple heuristics. To detect noisy samples of this type, an oracle model is required to calculate the semantic similarity between source and target sentence pairs.

In the last few years, there has been a growing interest in developing advanced parallel data filtering techniques, resulting in a shared task for "parallel corpus filtering" focusing on high and low resource language pairs (Koehn et al., 2019; Koehn et al., 2020). The majority of these approaches focus on rule-based pre-filtering followed by scoring with an oracle model trained on high-quality parallel data for the same language pairs (Esplà-Gomis et al., 2020). One of the early reported approaches with high accuracy (Junczys-Dowmunt, 2018) is based on predictor models that are applied in both forward and reverse directions. The models are trained on a given ”good quality” data and cross-entropy scores from both models are combined to calculate the quality of a given ”noisy” sentence pair. Such techniques rely on assumed ”good quality” data to train the oracle models which is not always available, especially in case of low-resource languages. On the other hand, there has been considerable research on learning “language agnostic” sentence representations (embeddings) which aim to produce equivalent representations of semantically equal sentences across languages. These models are trained on large monolingual and multilingual data and are shown to generate reasonable language independent representations, including for languages not included in their training data.

In this paper, we investigated the use of two well-known “language agnostic” sentence embedding models to asses training samples quality, namely: “Language-Agnostic Sentence Representations” (LASER) (Schwenk and Douze, 2017) and “Language-agnostic BERT Sentence Embedding” (LaBSE) (Feng et al., 2020a). We conducted experiments on the low resource language pairs used in the “Parallel Corpus filtering” shared tasks in WMT-19 and WMT-20. The performance of different filtering methods was evaluated by training MT models on size varying datasets drawn from the top ranked sentence pairs.

2 Related work

Some recent works have explored the use of cross-lingual representations for parallel corpus filtering. (Herold et al., 2021) experimented with various agreement scores to compute source-target sentence similarity based on word embeddings. Another work in the same line explored the use of multilingual BERT for filtering parallel corpora (Zhang et al., 2020). The authors leveraged the ability of the aforementioned model to project multilingual sentences into a shared space. Both these works can be considered the closest to ours. However, our work offers different contributions which can be listed as follows:

- They have mainly experimented with high resource languages, on the other hand, we focused on the low resource language pairs from the WMT “Parallel Corpus filtering” task namely: Khmer-English (KM-EN), Pashto-English (PS-EN), Sinhalese-English (SI-EN) and Nepali-English (NE-EN).
- Our work is centered around comparing the most well established models for generating language agnostic sentence representations while these other works experimented with either word embedding based approaches or models that are not explicitly optimized to align latent representations of parallel sentence pairs. Such models are proven to be inferior to the models that we experiment with in cross-lingual similarity tasks.
- To be able to mimic real world settings, we conduct large scale experiments by evaluating the best performing methods on a dataset that is complied by combining all the publicly available datasets for a given language pair which we collect from OPUS (Tiedemann, 2012). Furthermore, we complement our work by conducting human evaluations on chosen languages pairs.

3 Data filtering

In this paper, our aim is to explore the use of language agnostic sentence embedding models as oracles to determine the translation quality of parallel sentences in training datasets. For this purpose, we obtain language independent representations for each source and target sentence and determine the translation quality by calculating the cosine similarity between these embeddings. Sentence pairs are scored, ranked and then filtered based on a similarity threshold or by selecting a pre-defined percentage of the original data size. We experiment with two well-known sentence representation models: Language Agnostic Sentence Representations (*LASER*) and Language agnostic Bert sentence embeddings (*LaBSE*). We briefly explain these representations in the following subsections.

3.1 LASER

Language-Agnostic Sentence Representations (LASER) (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019) is based on exploiting neural machine translation architectures to learn join representations for different language pairs. They use separate encoders and decoders for each language and encourage bridging the representation gap using a multi-task learning framework with various training configurations. Their configurations vary based on the number of used encoders and decoders per batch update as follows: one-to-one, many-to-one, one-to-many, and many-to-many. Such training strategies are claimed to push the encoder representations of equivalent sentences in different languages closer to each other in the embedding space. At inference time, decoders are discarded and sentence embeddings are obtained by applying a max-pooling operation over the output of the encoder. LASER supports 93 languages belonging to 23 different scripts.

3.2 LaBSE

Language Agnostic Bert Sentence Embeddings (LaBSE) (Feng et al., 2020b), is based on combining some of the latest representation learning approaches namely: Masked Language Modeling (MLM) (Devlin et al., 2019), Translation Language Model (TLM) (Lample and Conneau, 2019), dual encoder translation ranking (Guo et al., 2018), and the use of additive margin softmax loss (Yang et al., 2019). Embeddings for source and target sentences were generated from a transformer model (Vaswani et al., 2017) pretrained using MLM and TML, they were then separately fed into a shared 12-layer transformer network. Finally, the output representations of parallel sentences were optimized to be similar to each other and distant from negative samples using an additive margin softmax loss. The latest LaBSE model supports 112 languages and is claimed to outperform LASER in parallel text retrieval.

4 Experimental setting

In WMT-19, the parallel corpus filtering task was organized for Sinhalese-English (SI-EN) and Nepali-English (NE-EN) and in WMT-20 the same task was organized for Khmer-English (KM-EN) and Pashto-English (PS-EN). In WMT-19, no scores were provided with the noisy training data to be used as a baseline system. In WMT-20, noisy training data was released along with the similarity scores computed using LASER embeddings to be used as the baselines.

We conducted two sets of experiments on the aforementioned language pairs. In the first set of experiments (see 4.1), we applied multiple filtering techniques on the noisy training datasets released by WMT parallel corpus filtering shared tasks, we then evaluated the performance of MT models trained on a varying number of the top ranked sentences according to the filtering systems as defined in the shared task descriptions. In the second set of experiments (see 4.2), we included additional training datasets that are publicly available and applied the approaches that resulted in the best results in our first set of experiments. We studied the performance of the MT systems trained on different data sub-samples to be able to determine an optimal similarity score threshold. The choice of the languages posed an interesting zero shot challenge because not all languages were used in training of both the models. As a result, not all four languages are supported by both the models. LASER only supports Khmer and Sinhalese, while LaBSE only supports Khmer, Nepali and Sinhalese. Pashto is not supported by either of the models. However it has been observed that both the models can be generalized well even to minority languages that are not supported by the models. This observation could possibly be attributed to transfer learning from languages in the training data which are closely related to these four languages.

	Year	Train	Valid	Test
KM-EN	WMT-20	4.1 m	2378	2320
PS-EN	WMT-20	1 m	3162	2719
SI-EN	WMT-19	3.3 m	2898	2766
NE-EN	WMT-19	2.2 m	2559	2835

Table 1: Parallel corpus filtering task WMT-19 and WMT-20 statistics. Data sizes are in number of sentence pairs.

4.1 WMT noisy data filtering

In this set of experiments, we aimed to replicate the setup of the "parallel corpus filtering" shared tasks in WMT-19 and WMT-20. Given a noisy corpus for a low resource language pair, the participating teams were required to submit quality scores for all the training samples. Filtering systems were then evaluated according to the performance of MT systems trained on varying size datasets (in terms of target tokens) sampled from the top ranked pairs according to the submitted quality scores. The MT systems were trained by the task organizers and used fixed model and hardware configurations to guarantee comparable assessment. Dataset statistics for each language pair are provided in Table 1

To establish a baseline, we train two MT systems on the entire data without any re-ranking or sub-sampling as follows:

- No filtering: The entire released noisy data is used for training without any filtering.

- Language ID filtering: Language filtering is applied to the training data using fastText toolkit (Joulin et al., 2016). Only the examples where the detected source and target language codes are not consistent with source and target languages are filtered out.

Sub-sampling experiments are setup as follows: sample sizes of [1, 2, 5] million target tokens are used for Nepali-English and Sinhalese-English and sample sizes of [2, 3, 5, 7] millions are used for Khmer-English and Pashto-English. Sub-sampling was performed using the script provided by the WMT organizers. The following filtering methods configurations are reported:

- Language ID filtering + LASER scoring: after applying language ID filtering, the remaining sentence pairs were scored and ranked according to the cosine similarity of the LASER embeddings.

- Language ID filtering + LaBSE scoring: same as the previous point but used LaBSE embeddings instead of LASER embeddings.

- Best performing systems in WMT tasks: To compare LASER and LaBSE filtering with the state-of-the-art language specific techniques, we selected AFRL (Erdmann and Gwinnup, 2019) for WMT-19 languages and Huawei (Açarçiçek et al., 2020) for WMT-20 languages. These models were chosen since they are the best performing models in the respective tasks where publicly available scores for the training samples are provided.[1]

4.2 Variable/mixed quality data filtering

Based on the first set of experiments as described in 4.1, we determined the best technique (out of LASER and LaBSE) for each language pair and applied it on a larger corpus that consists of samples of "unknown" quality. For each language pair, we first applied the standard language-id filtering, followed by scoring and re-ranking with the source-target embedding similarity score. We filtered the ranked corpus using different threshold values [0.5, 0.6, 0.7, 0.8, 0.9] of the similarity scores and trained NMT models on each filtered sub-sample. In this way we determined the best similarity threshold per language pair. We mainly used the datasets from OPUS repository and combined these datasets with the noisy corpus provided in the relevant WMT task. The details of the used datasets are provided in Table 2.

4.3 Training details

All the experiments in this paper were conducted using the same model architecture and training configuration. A TransformerBase model with the default configuration was trained using Opennmt-tf translation toolkit[2]. A Shared vocabulary of 10000 sub-words is trained using sentencepiece tokenizer(Kudo and Richardson, 2018). Models were trained for 1 million steps for the WMT data experiments and until convergence for the larger scale experiments. Convergence was defined as no significant change in the validation set performance according to BLEU scoring at 100,000 step increments.

Note that the NMT toolkit and the training configuration we used in this paper are different from those used in the WMT parallel corpus filtering tasks. This is because, in this paper, our purpose is not to do a direct comparison of a proposed methods with the results or methods reported in the WMT tasks but to empirically analyze utility of language agnostic embeddings for corpus filtering.

4.4 Evaluation metrics

BLEU score (Papineni et al., 2002) is the most commonly used automatic evaluation metric for machine translation performance. However, it has been recently criticized due its failure to correlate with human judgement. A recent study (Kocmi et al., 2021) conducted an extensive comparison of various MT evaluation metrics and found out that BLEU is inferior to other automatic metrics with respect to correlation to human judgements. They found that other metrics such as COMET (Rei et al., 2020) and ChrF (Popović, 2015) correlate much better. Therefore, following their recommendations, we reported model performances on COMET and ChrF in addition to BLEU. BLEU is calculated using sacreBLEU python implementation (Post, 2018).

ChrF is a character level n-gram F-score between generated translation and reference. Similar to BLEU, it calculates n-gram matches between

[1] Data can be downloaded from the websites of the respective WMT tasks

[2] https://github.com/OpenNMT/OpenNMT-tf

	Datasets	Sentences
KM-EN	CCAligned, GNOME, KDE4, Paracrawl, QED, wikimedia, XLEnt. WMT-20	4.8 M
PS-EN	CCAligned, GNOME, KDE4, Paracrawl, QED, wikimedia, XLEnt, WMT-20	1.4 M
SI-EN	CCAligned, CCMatrix, GNOME, KDE4, OpenSubtitles, Paracrawl, QED, Ubuntu, Wikimatrix, wikimedia, XLENT, WMT019	11 M
NE-EN	CCAligned, CCMatrix, GNOME, KDE4, OpenSubtitles, Paracrawl, QED, Ubuntu, Wikimatrix, wikimedia, XLENT, bible-uedin, WMT-19	2.1 M

Table 2: Combined data statistics for all language pairs for mixed data experiments.

translation and reference, however, at character level. ChrF is calculated using the same sacreBLEU implementation. For both BLEU and ChrF, statistical significance was measured using bootstrap re-sampling (Koehn, 2004) with 1000 samples. For the noisy-data-only experiments, we calculate statistical significance between LaBSE and LaSER based filtering as well as between best of language agnostic filtering with that of the competitor baseline as described in subsection 4.1. For the mixed-quality-data experiments, we compare the statistical significance between the no-filtering baseline with all other experiments (subsampling at various threshold values).

COMET is a neural network framework in which large pre-trained cross-lingual language models such as XLM-RoBERTa (Lample and Conneau, 2019) were fine-tuned on [source, hypothesis, reference] pairs in order to predict annotated human evaluation scores. We used a reference-based regression model which is built on top of XLM-R *wmt-comet-da*. This model covers all the languages in our study.

5 Results

5.1 Khmer-English

Table 3 shows the results for noisy data filtering experiment for KM-EN language pair. The performance of all models using the three evaluation metrics is monotonically consistent, i.e., higher performance with respect to one metric also means higher performance with respect to other metrics. The model achieves the lowest performance when trained on the entire data without any filtering. Filtering using language identification provides significant improvements. For sample sizes of 5 million and 7 million, sub-sampling based on LaBSE scoring performs the best, while for sample sizes of 2m and 3m, Huawei filtering (Açarçiçek et al., 2020) performs the best. This is consistent for both dev and test sets. Moreover, for all filtering methods, using a sample size of 7 million target tokens seems to perform the best, while using 2 million tokens seem to perform the worst.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 58 million					
NF	4.2	17.9	-1.17	4.8	20.1	-1.02
	target tokens = 15 million					
lg	6.3	24.8	-0.94	6.9	27.6	-0.85
	target tokens =7 million					
HW	**8.7**	32.1	-0.60	**10.2**	36.9	-0.44
LS	6.1	30.1	-0.75	7.0	33.1	-0.65
LB	8.5†	**32.2**†	**-0.59**	**10.2**†	**37.3**†	**-0.42**
	target tokens =5 million					
HW	8.0	32.4	-0.61	10.0	37.0	-0.45
LS	6.6	29.8	-0.75	7.2	32.8	-0.68
LB	**8.4***†	**32.5***	**-0.58**	**10.9***†	**38.1***	**-0.40**
	target tokens =3 million					
HW	**8.7***	**32.9***	**-0.58**	**10.5***	**38.0***	**-0.43**
LS	5.7	28.6	-0.82	6.5	32.2	-0.74
LB	8.2†	32.4†	-0.60	9.9†	37.2†	-0.45
	target tokens =2 million					
HW	**8.0***	**32.1***	**-0.63**	**9.6***	**36.9***	**-0.49**
LS	4.7	27.1	-0.90	5.4	30.4	-0.83
LB	7.3†	30.9†	-0.67	8.7†	35.3†	-0.55

Table 3: KM-EN WMT noisy data filtering. **NF**= No filtering, lg = language id filtering only, **HW**=Huawei system, **LS** = language id + LASER scoring, **LB**= language id + LaBSE scoring. Values in bold indicate the highest ranking system for each subsample category. * represents a statistically significant comparison between HW and best of the language agnostic method and † represents the same between LASER and LaBSE at $p < 0.01$.

Given that LaBSE performs significantly better than LASER when using the noisy data, we applied LaBSE scoring along with language filtering on a combined set of noisy and clean training data as described in 4.2. Table 4 describes the results for the mixed data experiments for KM-EN. Filtering only on language ID drops the performance on the development set when compared to using the full training data but the performance remained al-

	devtest			test		
	BLEU	ChrF++	COME	BLEU	ChrF	COME
NF	7.9	23.5	-0.97	8.7	27.5	-0.78
lg	**6.2**	25.5	-0.93	8.8	**30.7**	-0.70
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	**10.2**	**33.9**	-0.52	**12.1**	**39.1**	-0.33
0.6	10.2	33.8	-0.51	12.1	39.4	-0.33
0.7	10.0	33.8	-0.52	11.8	39.6	-0.32
0.8	9.9	34.3	-0.52	11.1	39.7	-0.32
0.9	.6	32.7	-0.60	9.8	37.5	-0.45

Table 4: KM-EN mixed data experiments. τ = LaBSE similarity score threshold. All results are statistically significance wrt no filtering baseline. Values in bold indicate training corresponding to highest score.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	10.3	34.5	-0.52	8.3	31.0	-0.67
lg	8.8	32.3	-0.60	6.7	28.9	-0.75
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	11.0	37.8	-0.39	9.8	36.0	-0.48
0.6	10.7	37.8	-0.40	9.8	36.1	-0.48
0.7	**11.2**	**38.3**	-0.37	**9.9**	**36.4**	-0.48
0.8	10.7	37.7	-0.42	9.5	35.9	-0.50
0.9	6.1	30.8	-0.73	5.8	29.8	-0.80

Table 6: PS-EN mixed data experiments. Legends have same meaning as Table 4.

most the same on the test set. However, when combined with LaBSE filtering, it provided significant improvements compared to no filtering at all. An embedding similarity score threshold of 0.6 seems to work the best on such dataset.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 12.9 million					
NF	7.7	31.6	-0.66	5.9	28.1	-0.80
	target tokens = 7.3 million					
lg	7.4	30.5	-0.67	5.7	27.2	-0.81
	target tokens = 7 million					
HW	**10.7**	**37.1**	**-0.43**	**8.8**	**34.2**	**-0.55**
LS	7.9	31.1	-0.65	5.8	27.7	-0.81
LB	8.2	31.7	-0.62	6.4	28.8	-0.75
	target tokens = 5 million					
HW	**10.2**	**37.3**	**-0.42**	**8.7**	**35.0**	**-0.52**
LS	7.3	31.6	-0.66	5.6	28.9	-0.78
LB	9.7	35.7	-0.47	8.0	33.2	-0.58
	target tokens = 3 million					
HW	**10.1***	**37.0***	**-0.43**	**9.3***	**35.4***	**-0.53**
LS	7.2	31.7	-0.68	6.0	30.2	-0.76
LB	**10.1**	**37.0**	-0.44	9.2	35.2	-0.54
	target tokens = 2 million					
HW	**9.3**	**35.9**	-0.49	**8.4**	**34.2**	-0.58
LS	6.4	31.3	-0.71	5.7	30.2	-0.77
LB	9.3†	35.5†	-0.51	8.3†	33.9†	-0.60

Table 5: PS-EN WMT noisy data filtering. Legends have same meaning as Table 3.

5.2 Pashto-English

Table 5 summarizes the results for noisy data experiments for PS-EN. Applying only language ID filtering causes some slight performance drop as compared to using the entire training data. Scoring and sub-sampling using LASER embeddings performs the worst for all sub-sampled sizes. For sample sizes of 7m and 5m, Huawei filtering technique performed the best while for sample sizes of 3m and 2m, both Huawei and LaBSE based filtering perform almost equally. The differences in performance between models are consistent across metrics and test sets. Regardless of the fact that Pashto is not supported by LaBSE, it's performance is comparable to the language specific filtering technique (Huawei).

LaBSE performed significantly better than laser on the noisy data experiments therefore, for the experiments with mixed quality data, we applied LaBSE based filtering. As observed in Table 6, for mixed data experiments, filtering only with the language ID seemed to drop the performance significantly. However, applying language ID filtering in combination with LaBSE based scoring with similarity thresholds in the range of [0.5, 0.8] provided substantial improvements as compared to using all the training data. However, with a score threshold of 0.9, the performance dropped even below that of only language ID filtering which implies that with a very high similarity threshold, a substantial amount of useful training samples get filtered out.

5.3 Sinhalese-English

Table 7 presents the results for Sinhalese-English noisy data filtering. An important observation for this language pair is the very low performance when no filtering is applied which might indicate high noise level in the crawled dataset. The performance drops further with language ID filtering. However, for this language pair, scoring with LaBSE outperforms both LASER as well as the best reported language specific approach (AFRL). The difference in scores is consistent across test sets as well as metrics. The best performance

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 45 million					
NF	3.8	21.7	-0.86	3.1	19.7	-0.91
	target tokens = 11 million					
lg	2.8	21.3	-0.91	2.0	19.9	-0.93
	target tokens =5 million					
AF	5.8	30.4	-0.53	5.2	29.5	-0.52
LS	**6.1***	**31.8***†	**-0.48**	**5.6***	**31.3***†	**-0.45**
LB	6.0†	31.3†	-0.50	5.4†	30.4†	-0.47
	target tokens =2 million					
AF	5.5	30.5	-0.57	5.0	29.8	-0.55
LS	5.7	32.0	-0.51	5.4	31.6	-0.50
LB	**7.3***†	**34.2***†	**-0.39**	**6.8***†	**33.7***†	**-0.37**
	target tokens =1 million					
AF	4.3	28.6	-0.64	4.0	28.3	-0.62
LS	3.3	27.4	-0.70	3.1	27.4	-0.67
LB	**6.4***†	**32.5***†	**-0.47**	**5.6***†	**31.6***†	**-0.46**

Table 7: SI-EN WMT noisy data filtering. **AF** = AFRL filtering. Other legends have same meaning as Table 3.

for the LaBSE model was observed with 2m samples which supports our hypothesis that the initial dataset is of lower quality than the other language pairs that we experimented with. Since LaBSE based filtering performed the best for the noisy data experiments, we applied it for mixed dataset filtering. Table 8 shows the mixed data filtering results for Sinhalese-English. The absolute scores using all metrics are substantially higher than those for the previous two language pairs. Simply applying language ID filtering provided significant improvements compared to using all the data without filtering. Further filtering using LaBSE provided additional improvements for all threshold values except for the threshold value of 0.9. The highest performance was observed when using a threshold score of 0.7.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	18.2	46.7	0.11	16.3	43.4	0.03
lg	19.2	47.3	0.14	16.5	43.7	0.05
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	19.9	49.0	0.21	18.8	48.2	0.22
0.6	**20.2**	**49.4**	0.22	**19.3**	**48.7**	0.23
0.7	20.2	49.5	0.23	19.0	48.6	0.23
0.8	19.5	49.0	0.21	18.5	48.2	0.22
0.9	15.8	45.6	0.06	14.9	44.7	0.06

Table 8: SI-EN mixed data experiments. Legends have same meaning as Table 4.

5.4 Nepali-English

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 35 million					
NF	0.7	13.2	-1.25	1.0	13.9	-1.17
	target tokens = 9 million					
lg	1.5	18.2	-1.07	1.8	18.9	-1.00
	5 million					
AF	**2.7***	**23.0***	**-0.85**	**2.8***	**24.5***	**-0.78**
LS	2.1	22.0	-0.92	2.7	23.8	-0.82
LB	2.4†	22.3†	-0.86	2.5†	23.5†	**-0.78**
	target tokens = 2 million					
AF	3.6	26.5	-0.78	3.8	28.5	-0.69
LS	2.4	24.2	-0.88	2.7	25.8	-0.78
LB	**5.2***†	**29.6***†	**-0.62**	**5.9***†	**31.6***†	**-0.54**
	target tokens = 1 million					
AF	2.7	25.4	-0.82	2.9	27.3	-0.75
LS	0.8	19.6	-1.03	1.2	20.7	-0.98
LB	**5.2***†	**29.4***†	**-0.64**	**6.1***†	**31.6***†	**-0.55**

Table 9: NE-EN WMT noisy data filtering. **AF** = AFRL filtering. **AF** = AFRL filtering. Other legends have same meaning as Table 3.

Table 9 shows the Nepali-English noisy data results. The absolute scores without filtering are the lowest when compared to other languages. Applying language ID filtering slightly improved the performance as compared to no filtering. LaBSE filtering performs significantly better that the other methods according to the majority of the evaluation metrics when using sample sizes of 1m and 2m samples. The language specific approach performs better than LaBSE on the 5m sub-sample however, the results on this sub-sample were the worst for all the approaches.

LaBSE based filtering results for the mixed quality dataset are presented in Table 10. The best performance was observed using a similarity threshold of 0.8. Consistent with the observations

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	9.1	30.1	-0.61	10.7	33.3	-0.50
lg	8.6	29.4	-0.63	10.3	32.6	-0.52
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	11.3	36.5	-0.33	12.8	39.9	-0.19
0.6	11.6	38.0	-0.27	14.1	42.2	-0.12
0.7	12.3	39.4	-0.22	14.7	43.4	-0.09
0.8	**12.8**	**40.4**	-0.19	**15.3**	**44.3**	-0.07
0.9	10.9	38.1	-0.32	12.9	41.8	-0.19

Table 10: NE-EN mixed data experiments. Legends have same meaning as Table 4.

for other language pairs, a very high threshold of 0.9 dropped the performance significantly as compared to other lower values.

5.5 Human evaluation

In order to further verify the certainty of model performances calculated using automatic scores, we additionally performed human evaluations for some experiments. Due to the low availability of human evaluators, we performed human evaluations only for Pashto-English and Sinhalese-English for WMT noisy data experiments corresponding to the results reported in Table 5 and 7 for the 5 million sub-sample task. For Pashto-English and Sinhalese-English, we randomly sampled 100 sentences from the development set which were rated by native speakers of the corresponding languages. The raters were directed to assign an integer adequacy score between [1,5] to each hypothesis translation (Koehn and Monz, 2006). The final average scores are shown in Table 11. For both language pairs, as expected, the reference translations scored the highest. For Pashto-English, the Huawei filtering method scored significantly higher than both LASER and LaBSE based filtering. For Sinhalese-English, while LASER model performed significantly lower, Huawei and LaBSE filtering performed approximately equally. These observations for both of the language pairs were consistent with the automatic evaluations in Table 5 and 7.

6 Discussion

In this paper, we presented an empirical evaluation of the use of language agnostic sentence representations to filter parallel data for low resource neural machine translation. Our experiments show that using similarity scores based on language agnostic embeddings to compute the quality of the sentence pairs performs competitively when compared to state-of-the-art language specific techniques for low resource languages.

Filtering out sentences based on automatic language detection seems to give inconsistent results, we think that this happens because of the accuracy differences of the used language detection tool across different languages. Further analysis needs to be done for better understanding.

Data filtering thresholds based on the similarity score or a pre-defined number of tokens seems to vary across languages and datasets. This can be attributed to two main factors namely: The inherent quality of the dataset and the performance of the cross-lingual embeddings when it comes to the language pair under evaluation. Further analysis needs to be conducted to understand the per language pair effects.

Based on our experiments, language agnostic approaches perform competitively and provide a simple and a hassle-free way of filtering parallel datasets. However, this isn't the case when the language pair is not supported by the cross-lingual embeddings models as shown in the PS-EN experiments. Further research is needed to develop and test approaches for incremental language addition to the cross-lingual embedding based models.

References

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Açarçiçek, Haluk, Talha Çolakoğlu, pınar ece aktan hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online, November. Association for Computational Linguistics.

Barbu, Eduard and Verginica Barbu Mititelu. 2018. A hybrid pipeline of rules and machine learning to filter web-crawled parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 867–871, Belgium, Brussels, October. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Erdmann, Grant and Jeremy Gwinnup. 2019. Quality and coverage: The afrl submission to the wmt19 parallel corpus filtering for low-resource conditions task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 269–272, Florence, Italy, August. Association for Computational Linguistics.

Esplà-Gomis, Miquel, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of*

	Reference	Huawei/AFRL	LASER	LaBSE
Pashto-English	4.4	2.4	1.7	1.8
Sinhalese-English	4.6	2.26	1.88	2.23

Table 11: Human evaluation average scores for Pashto-English and Sinhalese-English

the Fifth Conference on Machine Translation, pages 952–958, Online, November. Association for Computational Linguistics.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020a. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020b. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Guo, Mandy, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Herold, Christian, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2021. Data filtering using cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–172, Online, June. Association for Computational Linguistics.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Junczys-Dowmunt, Marcin. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 901–908, Belgium, Brussels, October. Association for Computational Linguistics.

Khayrallah, Huda and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia, July. Association for Computational Linguistics.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online, November. Association for Computational Linguistics.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Koehn, Philipp, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy, August. Association for Computational Linguistics.

Koehn, Philipp, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco GuzmÃ¡n. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online, November. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, Eduardo and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. ACL '02, page 311–318, USA. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Schwenk, Holger and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, August. Association for Computational Linguistics.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Yinfei, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In Kraus, Sarit, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Zhang, Boliang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online, July. Association for Computational Linguistics.

Construction of Japanese BERT with Fixed Token Embeddings

Arata Suganami , Hiroyuki Shinnou
Ibaraki University, Department of Computer and Information Sciences
4-12-1 Nakanarusawa, Hitachi, Ibaraki, Japan 316-8511
{22nm721n, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

Abstract

In this study, we propose a method to construct Japanese BERT using fixed Token Embedding to reduce BERT model construction time. This method constructs word embeddings in advance using word2vec and the corpus, which is used in learning BERT models. The token embeddings in the BERT model are set to the constructed word embeddings. The parameters of token embeddings are fixed, that is, not learned during BERT model learning. Therefore, the proposed method reduces the processing time of one epoch in BERT model training and obtains the BERT model in fewer epochs. In the experiments, we construct 1024-dimensional 4-layer and 128-dimensional 24-layer Japanese BERT model using conventional and proposed methods, respectively, and the effectiveness of the proposed method was verified.

1 Introduction

BERT (Devlin et al., 2019) is a high-performance pre-trained model, but its large model size requires significant time and computational resources to build. Another problem is the increase in model construction time and cost because of the huge data sets and models that come with the higher performance of pre-trained models (Sharir et al., 2020). To reduce the construction time of BERT, we propose a method to construct Japanese BERT by fixing the word distribution representation. Specifically, we constructed Japanese BERT by learning word embeddings in advance using word2vec (Mikolov et al., 2013) and fixing them as the Token Embedding of BERT. This reduces the time required to learn word embeddings during BERT construction.

In the experiments, we constructed 128-dimensional 24-layer and 1024-dimensional 4-layer Japanese BERTs using conventional and proposed methods, respectively, and the effectiveness of the proposed method was verified by comparing the model construction time and accuracy in a document classification task for Japanese news articles. The results showed that the proposed method reduced the construction time for both BERTs. The results of the document classification task showed that the proposed method can learn a model with the same level of accuracy using fewer epochs.

2 Related Work

2.1 Small size BERT

BERT is a huge model, and its training requires a significant amount of computation time and computer resources. The amount of time required to process inference using BERT is also significant. To address this problem, many studies have been conducted to miniaturize BERT.

DistilBERT (Sanh et al., 2019) is a miniaturized BERT built using distillation (Hinton et al., 2019).

ALBERT (Lan et al., 2019) reduced the number of parameters of BERT using two methods, such as Cross-Layer Parameter Sharing and Factorized Embedding Parameterization. Cross-Layer Parameter Sharing is a method to used for sharing the parameters of each layer of BERT. Factorized Embedding Parameterization decomposes a matrix of size V×H representing the word embedding into a product of a matrix of size V×E and a matrix of size E×H, where E is a small number (e.g. 128), the number of parameters in BERT's word embedding representation can be reduced by approximately 13%.

TinyBERT (Jiao et al., 2020) is a miniaturized version of BERT that uses a two-step distillation process. The first distillation stage is specialized for transformers. The transformer is classified into four, and a loss function is set for each layer. The BERT fine-tuned in the downstream task is used as the teacher model, in the second stage of distillation, while the TinyBERT constructed in the first stage is used as the student model. To perform Transformer-specific distillation, the training data is augmented with task-specific data.

Q8BERT (Zafrir et al., 2019) is a miniaturized version of BERT that quantized the weights of all coupling and embedding layers to 8-bits. However, operations that require high accuracy (softmax and layer normalization) are still 32-bit. In learning, the forward process uses the quantized values of the parameters, while the back backpropagation process uses the pre-quantized values, a technique known as Fake Quantization.

MobileBERT (Sun et al., 2020) is a small and lightweight BERT that can run on small devices. While most miniaturized BERTs are miniaturized by reducing the number of layers, MobileBERT is miniaturized by reducing the dimensionality of the embedding representation. Therefore, BERT-large, which has more layers and a higher dimensionality of the embedding representation than BERT, is targeted for miniaturization, and a Bottleneck layer and an inverse Bottleneck layer are added to the model to match the dimensionality of the embedding representation. The model is trained using Progressive Knowledge Transfer, a method in which training is performed sequentially, starting from the bottom layer.

Poor Man's BERT (Sajjad et al., 2020) is a simple BERT with some layers removed. Fine-tuning can achieve a certain level of performance by simply adjusting the layers to be removed according to the task.

The method proposed in this study does not learn BERT token embeddings. Therefore, the number of parameters to be learned in the BERT model is reduced, which can be regarded as miniaturization of the BERT model.

2.2 Learning method of BERT

The BERT model is primarily trained using MLM (Masked Language Model), but the model building time can be reduced by improving the training method. In PMI-Masking (Levine et al., 2021), a vocabulary set is set designed based on improved PMI (Pointwise Mutual Information) to facilitate the prediction of Mask words in MLM. We demonstrated how this learning strategy allows us to construct a high-performing BERT model with fewer epochs.

EarlyBERT (Chen et al., 2021) uses the Lottery Ticket Hypothesis (Frankle and Carbin, 2019) methodology used in computer vision tasks to identify structured winners in the early stages of BERT training result in a model with the same performance as conventional BERT ranging form 35% to 45% less cost.

The proposed method in this atudy learns BERT token embeddings from word2vec, with the token embeddings not having to be obtained by training the entire BERT model. The processing time for word2vec training is significantly less than that for BERT training.

2.3 Task specific BERT

TLM (Task-driven Language Modeling) (Yao et al., 2021) is a method used to learn BERT from scratch using only task-related data. The time required to build a model is 1/10-1/100 of that required for existing methods, and the performance of the built model outperforms that of general BERT models. Furthermore, the constructed models outperformed DAPT-TAPT (Gururangan et al., 2020), which is an excellent method for domain adaptation using BERT.

The method proposed in this study can be used task-specifically and directly in conjunction with TLM.

3 Method

In this study, we propose to fix the word embedding to reduce the construction time of BERT. The following procedures were used to

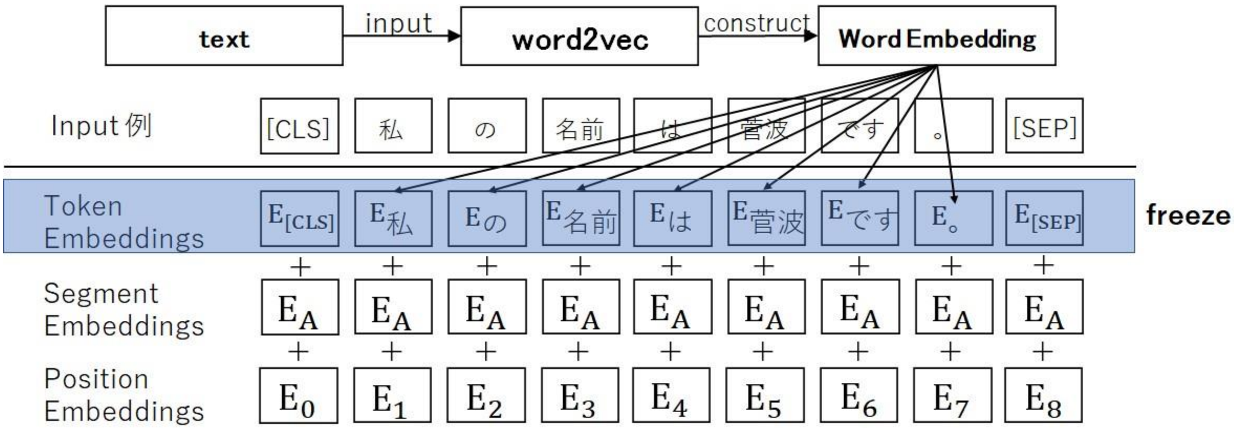


Figure 1: Fixing Word Distributed Representations

achieve this. Figure 1 shows an image of the proposed method.

1. Construct word embeddings from the text using word2vec. This gives word embeddings for all words in the text.
2. Match the id of the word used for BERT training with the id of the word in the word embedding learned with word2vec.
3. When constructing BERT, the word embedding constructed in 1, is fixed as Token Embedding, and the parameters of Token Embedding are frozen so that they are not updated.

Word2vec fixes the parameters of Token Embedding to the learned word embedding by freezing them. The number of parameters can be reduced by 128 × 32,000 = 4,096,000 for a 128-dimensional 24-layer BERT and 1024 × 32,000 = 32,768,000 for a 1024-dimensional 4-layer BERT because the vocabulary size of the BERT is set to 32,000 in this study. The time required to learn BERT can be omitted, resulting in a shorter construction time.

4 Method

4.1 Construction of Word Embeddings

As the corpus for pre-training, we used dumped data from the Japanese Wikipedia as of November 1. These data are from Japanese Wikipedia, containing 1.17 billion characters with a data size of 3.26 GB. This data was text cleaned using a wikiextractor, the text was preprocessed, and multiple txt files were combined into a single text. The completed text was split into words using the "cl-tohoku/bert-base-japanese" tokenizer in Tohoku University's BERT. Then, we constructed a 128-dimensional word embedding and a 1024-dimensional word embedding from the text. using the Python library “gensim” word2vec. The learning algorithm used in this study was CBOW.

4.2 Construction of Japanese BERT

We built a 128-dimensional 24-layer Japanese BERT and a 1024-dimensional 4-layer Japanese BERT for 8 epochs each using a program for building Japanese BERT published on GitHub. Additionally, we modified the program to allow the construction of Japanese BERT by fixing the word variance representation constructed in section 4.1 to Token Embedding, and we used the proposed method to construct a 128-dimensional 24-layer Japanese BERT and a 1024-dimensional 4-layer Japanese BERT for 8 epochs each. The four Japanese BERTs constructed in this manner have a maximum input word count of 256 and a vocabulary of 32,000. The model building time was recorded for each epoch of model building. The flow of the Japanese BERT construction described so far is shown in Figure 2.

4.3 Model Evaluation

After constructing the Japanese BERT, we evaluated the model by comparing the percentage of correct answers in the document classification task. Livedoor news corpus published by Lonwit, Inc. was used for the evaluation data. The 7376

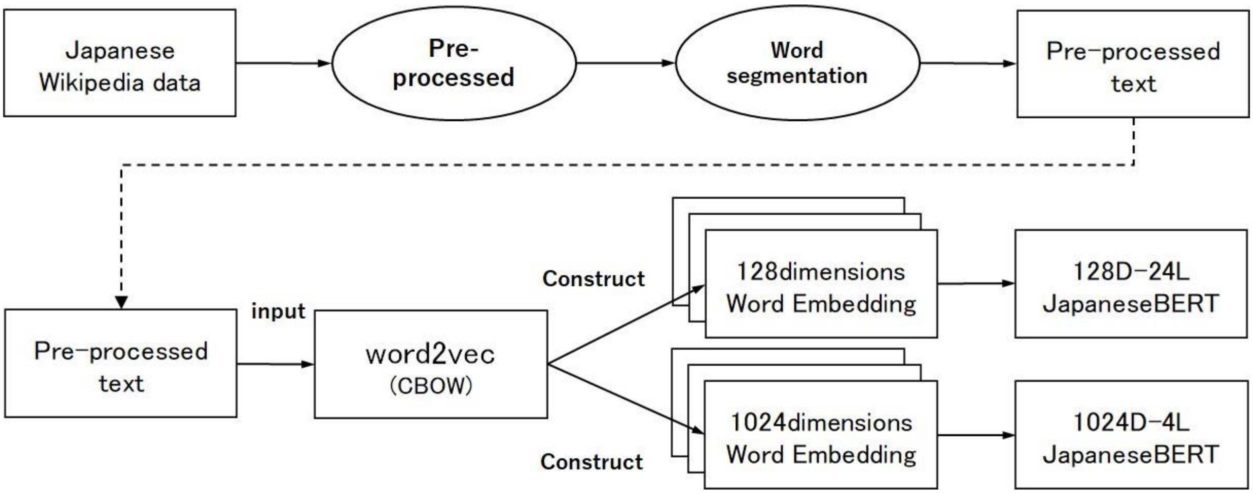


Figure 2: Flow chart for building Japanese BERT

articles in the corpus were assigned labels from 0 to 8 based on the category of the article, and the 7376 articles were assigned to the training data, validation data, and test data (Table 1).

label	category	train	valid	test
0	Dokujo tsushin	87	87	697
1	IT lifehack	87	87	697
2	Kaji Channel	86	86	693
3	Livedoor HOMME	51	51	410
4	MOVIE ENTER	87	87	697
5	Peachy	84	84	675
6	Smax	87	87	697
7	Sports Watch	90	90	721
8	Topic News	77	77	617
Total		736	736	5904

Table 1: Breakdown of data used

The model was trained using the training data, and the percentage of correct answers for the trained model was measured using the validation data after each training session. The loop of training and validation was set to a maximum of 15 epochs (Figure 3). When the highest percentage of correct answers to the validation data was not updated five times, the model training was terminated as Early Stopping, and the model with the highest percentage of correct answers up to that point was saved as the best model. Finally, the best model was subjected to document classification using test data to predict the category to which an article belongs based on the text of the article, and the percentage of correct answers was measured.

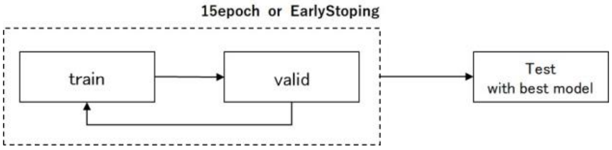


Figure 3: Model Evaluation

5 Result

5.1 Construction time

Table 2 shows the time per epoch taken to build the Japanese BERT. The "H" in the model name represents the dimensionality of the word embedding, and the "L" represents the number of layers. "N" indicates that the construction was performed using the conventional method, whereas "W" indicates that the construction was performed using the proposed method. The proposed method reduces construction time by 4 min per epoch for 24 layers of 128 dimensions and by 27 min per epoch for 4 layers of 1024 dimensions.

Model	Construction Time
H128-L24-N	16h49m
H128-L24-W	16h45m
H1024-L4-N	18h05m
H1024-L4-W	17h38m

Table 2: Construction time per epoch

5.2 Document Classification Accuracy

Table 3 shows the percentage of correct answers in the document classification task performed using the constructed Japanese BERT. The best correct response rate is indicated by bolded numbers. The proposed method did not improve the accuracy of the 24 layers with 128 dimensions, but it improved the accuracy of the 4 layers with 1024 dimensions.

6 Result

6.1 Model building Time

Table 2 shows that the proposed method reduces the construction time by 4 min for 24 layers of 128 dimensions and by 27 min for 4 layers of 1024 dimensions, indicating a reduction in construction time due to the omission of learning word embeddings in BERT. The fact that the time reduction was greater in the 1024-dimensional 4-layer case indicates that the larger the dimensionality of the Japanese BERT, the greater the number of parameters that can be reduced, thus the greater the time reduction by the proposed method.

6.2 Model Accuracy

Table 3 shows that the proposed method improved the accuracy of the 1024-dimensional, 4-layer model. The best scores were also recorded early by the proposed method for both BERTs. Therefore, we assumed that the accuracy of Japanese BERT fixed with the word embedding created using word2vec will converge at earlier epochs. If accuracy can be achieved at an early stage while improving accuracy in this way, it is possible to build a highly accurate model in fewer epochs when constructing BERT compared to the conventional method, therefore, the construction time may be significantly reduced.

7 Conclusion

To reduce the construction time of BERT, this study focuses on the time required to learn word embedding during BERT construction, and proposes a method to construct Japanese BERT by fixing a previously learned word embedding using word2vec to Token Embedding. In the experiment, we constructed a 128-dimensional 24-layer and a 1024-dimensional 4-layer Japanese BERT and measured the construction time, and found that the construction time could be reduced. The accuracy measurements in a document classification task showed that accuracy improved and converged at a lower number of epochs. This shows that fixing the word embedding can reduce the construction time of BERT.

	H128 L24-N	H128 L24-W	H1024 L4-N	H1024 L4-W
ep1	0.7033	0.6982	0.6386	**0.7236**
ep2	0.7100	0.7060	0.6811	0.7107
ep3	0.6922	**0.7148**	0.6765	0.6816
ep4	0.6865	0.6738	0.6822	0.6685
ep5	0.7122	0.6963	0.6811	0.6992
ep6	0.6729	0.6975	0.7041	0.7002
ep7	0.7029	0.6897	**0.7043**	0.6753
ep8	**0.7204**	0.7126	0.6941	0.6692

Table 3: Document Classification Accuracy

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. International Conference on Learning Representations.

Hinton, Geoffrey and Vinyals, Oriol and Dean, Jeff. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Yoav Levine and Barak Lenz and Opher Lieber and Omri Abend and Kevin Leyton-Brown and Moshe Tennenholtz and Yoav Shoham. 2021. {PMI}-Masking: Principled masking of correlated spans. International Conference on Learning Representations.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ICLR 2013.

Zafrir, Ofir and Boudoukh, Guy and Izsak, Peter and Wasserblat, Moshe. 2019. Q8bert: Quantized 8bit bert. arXiv preprint arXiv:1910.06188.

Hassan Sajjad and Fahim Dalvi and Nadir Durrani and Preslav Nakov. 2020. Poor Man’s BERT: Smaller and Faster Transformer Models. CoRR, abs/2004.03844.

Or Sharir, Barak Peleg, Yoav Shoham. 2020. The Cost of Training NLP Models : A Concise Overview CoRR abs/2004.08900.

Gururangan, Suchin and Marasovic, Ana and Swayamdipta, Swabha and Lo, Kyle and Beltagy, Iz and Downey, Doug and Smith, Noah A. 2020. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Association for Computational Linguistics.

Sanh, Victor and Debut, Lysandre and Chaumond, Julien and Wolf, Thomas. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Chen, Xiaohan and Cheng, Yu and Wang, Shuohang and Gan, Zhe and Wang, Zhangyang and Liu, Jingjing. 2021. EarlyBERT: Efficient BERT Training via Earlybird Lottery Tickets. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2195–2207, Association for Computational Linguistics.

Jiao, Xiaoqi and Yin, Yichun and Shang, Lifeng and Jiang, Xin and Chen, Xiao and Li, Linlin and Wang, Fang and Liu, Qun. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. Association for Computational Linguistics, pages 4163–4174.

Yao, Xingcheng and Zheng, Yanan and Yang, Xiaocong and Yang, Zhilin. 2021. NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework. International Conference on Learning Representations.

Lan, Zhenzhong and Chen, Mingda and Goodman, Sebastian and Gimpel, Kevin and Sharma, Piyush and Soricut, Radu. 2019. Albert: A lite bert for selfsupervised learning of language representations. arXiv preprint arXiv:1909.11942.

Sun, Zhiqing and Yu, Hongkun and Song, Xiaodan and Liu, Renjie and Yang, Yiming and Zhou, Denny. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2158–2170, Association for Computational Linguistics.

Crosslinguistic Influence on VOT Spectrum: A Comparative Study on English, Mandarin and Min

Jarry Chia-Wei Chuang
Department of English
National Chengchi University
Taipei, Taiwan
cwchuang.academia@gmail.com
ORCID: 0000-0002-3029-4463

Abstract

Crosslinguistic comparison of VOT has indicated linguistic transfer of voicing and aspiration contrasts in many languages. Mandarin has clear aspiration contrasts for voiceless stops, while Min presents another complicated VOT pattern, where voicing and aspiration contrasts are involved. The present study makes a crosslinguistic comparison between languages with voicing and aspiration contrasts as well as the potential linguistic transfer of VOT in English contexts from Mandarin and Min. There are three subject groups, including American English natives and Mandarin-Min bilinguals with different levels of Min-fluency. Mandarin-Min bilinguals have more aspirations and higher VOTs for aspirated voiceless stops than English natives. They also present two surfaces for English underlying voiced stops, voiced and unaspirated voiceless. Different levels of Min fluency are found to influence the tendencies towards voiced or unaspirated voiceless representations of English voiced stops. The overall finding presents a clear crosslinguistic influence on VOT patterns.

Keywords: Voice onset time (VOT), crosslinguistic, English, Mandarin, Min.

1 Introduction

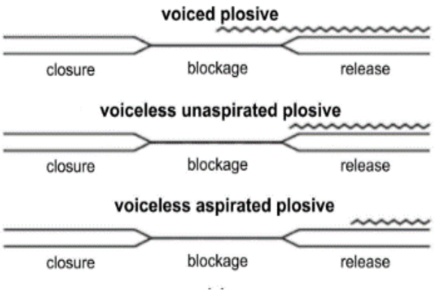
Researchers have been long explored phonetic and phonological acquisition in Mandarin-speaking ESL and EFL contexts (Chao & Chen, 2008; Chuang, 2021; Liu, 2017; among others). In Taiwan, Mandarin is the dominant language and English is the first foreign language, which has been taught from primary education to tertiary education. Taiwan EFL learners have been reported to have several phenomena of linguistic transfer from Mandarin to English, one of which is concerned with voicing contrast in stops. Previous studies on crosslinguistic influence have only taken Mandarin as the primary variable (Chao & Chen, 2008), while Crosslinguistic influences should be carefully noted. Taiwan is a multilingual community and most of the Taiwanese people are at least bilingual, though their L2 fluency may, subtly or divergently, differ from one to another. In Taiwan they are likely to acquire Min, Haka, or Austronesian languages; among all, Mandarin-Min bilinguals is the majority. To better capture the linguistic transfer of VOT, the paper will take different linguistic backgrounds and experiences (especially Mandarin-Min bilingualism) into account, revisiting crosslinguistic influences on VOTs in Taiwan English contexts via acoustic analysis.

2 Literature Review

2.1 VOT

Voice onset time (VOT) has led to phonetic studies in a vast number of languages. VOT marks the release of obstruction as well as reflects laryngeal vibration. Though its reliability for voicing distinction has been doubted in intervocalic or word-final positions (Docherty, 2011), it is still convincing that VOT can help identify voicing contrast in syllable-initial positions. VOTs show a categorical pattern for voicing and aspiration contrasts, in which VOT patterns can be classified into three possibilities (Lisker & Abramson, 1964), as shown in **Figure 1**.

Figure 1. Phonetic VOT patterns of plosives



The three-way distinction includes: (1) negative VOT (long lead), in which vocal folds vibrate far prior to the release of obstruction; (2) zero VOT (short lag), in which laryngeal vibration almost coincides with the unblocking of obstruction, often briefly delayed; (3) positive VOT (long lag), of which the occurrence is based on time priority of unobstructed airflow over the vibration. The three-way VOT patterns correspond to the voicing and aspiration of, particularly, stop consonants: (1) negative VOT for voiced stops, (2) zero VOT for voiceless unaspirated stops, and (3) positive VOT for voiceless aspirated stops.

The categorization is also be applied to the phonological analysis. [±voice] and [±spread glottis] feature in the linear phonological framework. Under Optimality Theory (OT), constraints in markedness and faithfulness adopt dichotomized judgments in voicing and aspiration as well.

2.2 VOT Spectrum

As the categorization seems well constructed for VOT, crosslinguistic findings reveal the deficiencies of the three-way VOT categorization and of the binary distinction in voicing and aspiration. In bilingual/multilingual contexts, speakers can produce divergent VOT values in the same category. To well account for the crosslinguistic evidence, Cho and Ladefoged (1999) proposed that VOT patterns should be better presented in a spectrum.

Aside from crosslinguistic influences on VOT, place of articulation and vowel contexts have been proven influential for VOT values in the same category. According to the aerodynamics resulting from jaw movements, the production of velar stops may contract the supraglottal cavity. It requires longer VOT, as it recovers from the formation of obstruction (Cho & Ladefoged, 1999). As for the surrounding vowels, tenseness and height of vowels might also contribute to the increase in VOT values (Klatt, 1975; Port & Rotunno, 1979; Weismer, 1979). So, a fine inspection of VOT values is thus required when we explore the voicing and aspiration contrasts, which has been accordingly considered in the experiment design.

2.3 Phonological & Phonetic Comparison

Normally, VOT patterns have corresponding phonological inventories of word-initial stops in the underlying representation (UR). Voicing and aspiration contrasts in English, Mandarin, and Min are differently distributed in UR. Crosslinguistic consonantal distributions of voicing and aspiration lead to a phonological comparison in **Table 1**.

	English	**Mandarin**	**Min**
Aspirated Voiceless		/p^h, t^h, k^h/	/p^h, t^h, k^h/
Unaspirated Voiceless	/p, t, k/	/p, t, k/	/p, t, k/
Voiced	/b, d, g/		/b, d, g/

Table 1. Phonological representation of word-initial contrasts in voicing and aspiration

Phonological representation may not be the final output, for aspiration and phonetic realization rule, as illustrated in **Figure 2.** Word-initial stops in English can phonologically be voiced /b, d, g/ or voiceless /p, t, k/ in UR. In SR, voiceless stops in word-initials become aspirated. Besides, underlying voiced stops in English have been reported to have two representations in SR, on the basis of the phonetic realization rule. One of them is voiced stops with vibration delays in VOTs, and the other is unaspirated voiceless stops with 15-20-ms vibration delays in VOTs. Among two surfaces of English voiced stops, the latter is the majority and greatly controls the mean VOT value. Phonetically speaking, English stops in word-initials mostly present phonological voicing contrast by aspiration in SR (i.e., phonetic difference between Zero VOTs and Positive VOTs).

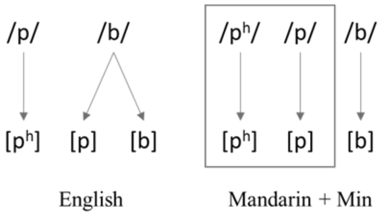


Figure 2. Phonetic realization of word-initial stops

In Mandarin and Min, aspiration contrasts of voiceless stops phonetically and phonologically play the major function of VOT. Mandarin has voiceless (un-)aspirated stops only, so there are no voicing contrasts. As for Min, the phonological inventory is sophisticated, with three categorized voicing and aspiration contrast phototactically acceptable in word-initial positions. Phonological patterns of onset plosives vary between English, Mandarin and Min, in which crosslinguistic performances are expected to display a wide array of patterns.

2.4 VOTs of Mandarin and English

In Chao and Chen's (2008) study, Mandarin-English VOT patterns have been comparatively examined. Both Mandarin and English VOTs fit the dichotomized classification (short lag vs. long lag), with aspiration contrasts in outputs only. Further visiting the voiceless aspirated stops, they indicate that VOTs in two languages are unidentical, in which VOTs of [p^h, t^h, k^h] in Mandarin are higher than those in English, which also provides evidence that VOT is better presented in a spectrum rather than a three-way distinction. Given the results, it is intriguing to see if crosslinguistic influence can trigger a linguistic transfer of VOTs from Mandarin to English.

3 Method

3.1 Subjects

Subjects were limited to be English natives or Mandarin-Min Bilinguals, with their ages ranging from 18 to 31 (*M* = 21.35; *SD* = 3.26).; in total, there have been 11 American English natives (5M; 6F) and 31 Mandarin-Min bilingual speakers (16M; 15F) in participation with the experiment. All the Mandarin-Min bilinguals were Taiwan Mandarin natives. Their Min fluency has been preliminarily self-identified and secondarily validified by 3 linguistically-trained Min natives in a 5-minute speaking test. Their Min fluency was scored with a Likert scale, from 1 (low) to 5(high) (M_L = 1.76; M_H = 4.67). Participants would be excluded as Min fluency was around 3 (intermediate) or presented a distinct mismatch between self-identified results and speaking checks by Min natives. There were 17 subjects with low fluency in Min (8M; 9F) and 14 subjects being fluent in both Mandarin and Min (7M; 7F). No speech disorders or diseases have been found.

Mandarin-Min transfer should be particularly noted here. Mandarin is much more dominant in Taiwan than Min. Here, we define Mandarin as their first language and Min as their second language with intimate language contact. Most of the linguistic transfers are Mandarin-to-Min, while the increase of Min fluency in Mandarin-Min bilinguals may lessen the transfer. High Min-fluency speakers may also have more linguistic habits brought from Min to other languages.

3.2 Stimuli

Word-initial stops were the major focus of the present study, as shown in **Table 1**. A total of 60 English words began with voiced and aspirated voiceless plosives and were adopted as the experimental materials for all the subjects. Mandarin-Min bilinguals were asked to read out extra 60 Mandarin words and 90 Min words, which were designed as disyllabic words for disambiguation of single-word senses. Tokens would be 10 words per stop in a language. Vowel contexts followed by target stops have been set with 5 tense high vowels and 5 lax low vowels for a stop. In total, 9 American English natives produced 540 English tokens. As for Mandarin-Min bilinguals, low Min fluency subjects produced 2550 tokens high Min fluency subjects produced 2100 tokens. It should be noted that 3 English tokens, 2 Mandarin tokens, and 7 Min tokens were reported to be invalid and have been excluded, with neglectable data loss.

3.3 Procedure

Before participating in the experiment, subjects had a self-evaluation of Min fluency and advanced evaluation by Min natives. All the subjects were then asked to read out English tokens in random order. Mandarin-Min bilinguals would further read Mandarin and Min tokens to check the VOT patterns in bilingual conditions.

3.4 Data Analysis

Acoustic data from the experiment has been imported to Praat 6.1.50 (Boersma, 2006) for analysis. VOTs were measured with waveforms and spectrograms, based on the interval between the release burst and the glottal vibration, as marked in **Figure 3**. The red is marked for the measurement of VOT.

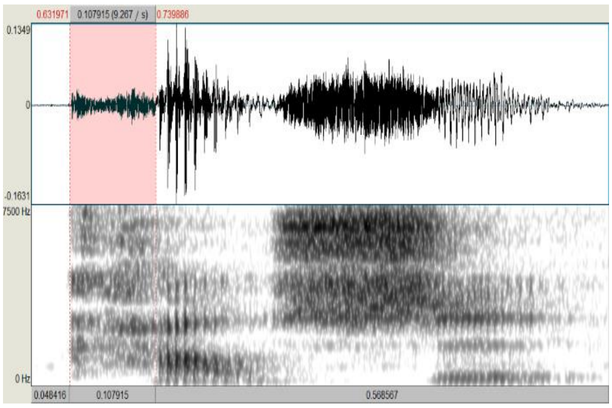


Figure 3. Acoustic analysis of *kǎo shì* ‘exam’ in the waveform and spectrograph.

4 Results & Discussion

4.1 Mean VOT

VOT patterns in Mandarin, Min and English contexts are presented as follows, along with the comparison between subjects with low (L) and high (H) Min fluency as well as English natives (E).

4.1.1 VOTs in Mandarin

For Mandarin VOT patterns, 3 aspirated voiceless stops /p^h, t^h, k^h/ and 3 unaspirated voiceless stops /p, t, k/ have been examined in **Table 2**.

	/p^h/	/t^h/	/k^h/	/p/	/t/	/k/
L	87.6	84.1	93.2	16.3	13.6	25.4
H	89.2	85.7	97.3	11.9	16.9	28.7
L+H	88.3	84.8	95.1	14.3	15.1	26.9

Table 2. Mean VOT values of Mandarin-Min bilinguals in Mandarin contexts

Mean VOT values show stops in Mandarin contexts contribute more to positive VOTs. We figure out no distinct divergence of VOT values between subjects with low and high Min frequency. In average, VOT values of aspirated voiceless stops are around 90 [$\bar{X}_1$(p^h, L+H) = 88.3; $\bar{X}_1$ (t^h, L+H) = 84.8; $\bar{X}_1$ (k^h, L+H) = 95.1]. As for unaspirated voiceless stops, the average VOT values are positive as well [$\bar{X}_1$ (p, L+H) = 14.3; $\bar{X}_1$ (t, L+H) = 15.1; $\bar{X}_1$ (k, L+H) = 26.9]. Mandarin VOT patterns generally show Mandarin aspirated voiceless stops have strong aspiration, with a long delay of glottal vibration.

4.1.2 VOTs in Min

Min phonology permits three major kinds of word-initial distributions in the VOT spectrum, including 3 aspirated voiceless stops /p^h, t^h, k^h/, 3 unaspirated voiceless stops /p, t, k/, and 3 voiced stops /b, d, g/. VOT patterns in Min contexts are shown in **Table 3**.

	/p^h/	/t^h/	/k^h/	/p/	/t/	/k/
L	99.6	97.1	101.7	11.7	10.6	23.2
H	102.3	103.8	111.9	8.2	9.3	27.1
L+H	100.8	100.1	106.3	10.1	10.0	25.0

	/b/	/d/	/g/
L	-97.1	-92.4	-133.4
H	-134.7	-119.5	-157.9
L+H	-114.1	-104.6	-144.5

Table 3. Mean VOT values of Mandarin-Min bilinguals in Min contexts

In Min, aspirated voiceless stops have positive VOTs, unaspirated voiceless stops have nearly Zero VOTs, and voiced stops have negative VOTs. Mean VOTs of aspirated voiceless stops falls on around

100 ms [$\bar{X}_2$(p^h, L+H) =100.8; $\bar{X}_2$ (t^h, L+H) =100.1; $\bar{X}_2$ (k^h, L+H) =106.3], and those of unaspirated voiceless are around 10-25 [$\bar{X}_2$(p, L+H)=10.1; $\bar{X}_2$ (t, L+H)=10.0; $\bar{X}_2$ (k, , L+H)=25.0]. VOTs of voiced stops in Min show the burst releases are much earlier than glottal vibration [$\bar{X}_2$(b, L+H)= －114.1; $\bar{X}_2$ (d, L+H) = －104.6; $\bar{X}_2$ (g, L+H)= －144.5]. The VOT patterns of Min are distinguishable for the clear three-way distinction.

4.1.3 VOTs in English

English phonology has only voicing contrasts, while it phonetically allows aspirated voiceless stops for underlying voiceless stops and unaspirated voiceless stops for voiced stops to appear word-initially. Crosslinguistic VOT patterns offer a well comparison between phonological representations and phonetic realization, as presented in **Table 4**.

	/p/	/t/	/k/	/b/	/d/	/g/
E	49.5	55.7	77.3	6.1	8.1	20.7
				-52.2	-70.5	-66.3
L	89.6	92.0	99.3	14.1	13.7	23.7
				-100.54	-95.9	-111.8
H	99.8	99.5	109.4	15.5	14.3	19.7
				-122.5	-125.1	-130.6
L+H	94.2	95.4	103.9	14.8	14.1	21.7
				-115.4	-115.1	-121.9

Note: Word-initial /p, t, k/ are [p^h, t^h, k^h]; VOTs of /b, d, g/ can be presented in short lag or long lead for phonetic realization.

Table 4. Mean VOT values in English contexts

Regarding aspirated voiceless stops for underlying voiceless stops, VOTs of English natives have vibration delays of around 50-80 ms [$\bar{X}_3$(p, E)=49.5; $\bar{X}_3$ (t, E)=55.7; $\bar{X}_3$ (k, E)=77.3] and Mandarin-Min bilinguals present much longer VOTs, up to 90-110 ms in average [$\bar{X}_3$(p, L+H) =94.2; $\bar{X}_3$ (t, L+H) =95.4; $\bar{X}_3$ (k, L+H) =103.9].

As for voiced stops [b, d, g], acoustic data has presented an intricate pattern. English natives present Zero VOTs [$\bar{X}_3$(b, E_1)= 6.1; $\bar{X}_3$ (d, E_1)= 8.1; $\bar{X}_3$ (g, E_1)=20.7] as well as negative VOTs (in the minority) [$\bar{X}_3$(b, E_2)= －52.2; $\bar{X}_3$ (d, E_2)= －70.5; $\bar{X}_3$ (g, E_2)= － 66.3]. Besides, Mandarin-Min bilinguals also have two types of VOTs for English voiced stops: Their mean VOT values for English voiced stops are about 15 ~ 25 ms [$\bar{X}_3$(b, L+H_1)= 14.8; $\bar{X}_3$ (d, L+H_1)= －14.1; $\bar{X}_3$ (g, L+H_1)= 21.7], and around－115 ~－120 ms [$\bar{X}$3(b, L+H_2)=－115.4; $\bar{X}$3 (d, L+H_2)=－115.1; $\bar{X}$3 (g, L+H_2)=－121.9]. Mandarin-Min bilinguals' VOT patterns of English voiced stops are more complicated than what figures show in the table, which will be further examined in 4.2.

4.2 VOT Distribution of English voiced stops

Though **Table 4** seems to show that Mandarin-Min bilinguals' potential tendencies towards negative VOTs as they produce English voiced stops, mean VOT values do not provide sufficient cues for such accounts. Their distributions are actually divergent, which needs careful analysis of the VOT distributions.

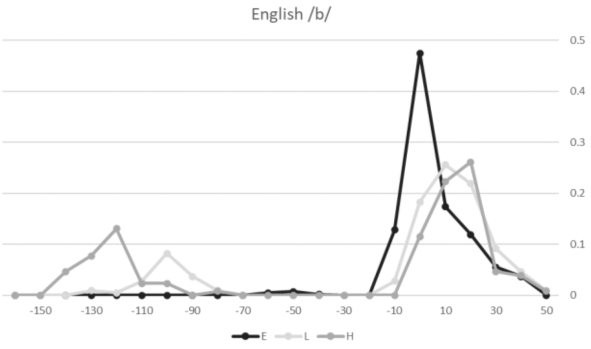


Figure 4. Surfaces of underlying /b/ in English

In **Figure 4**, English native speakers' VOT values for /b/ reach a peak at nearly 0 ms. They mostly produce Zero-VOT [p] for /b/. It should be noted that /b/ can [p] or [b] in the surface for phonetic realization rules, so a little number of subjects still present negative VOTs. In addition, Mandarin-Min bilinguals also produce /b/ in similar ways, but their occurrence rates show two divergent surface representations, which obviously differ from those produced by English natives. Mandarin-Min bilinguals' production of /b/ can show nearly Zero VOT (40 ~－10 ms) as well as negative VOTs centering on around －80 ~－150 ms, in which subjects with different Min fluency show the negative VOT peaks differently. High Min-Fluency subjects produce an earlier peak of negative VOTs (－120 ms) than low Min-fluency subjects (－100 ms), which reaches a statical significance ($p < 0.05$).

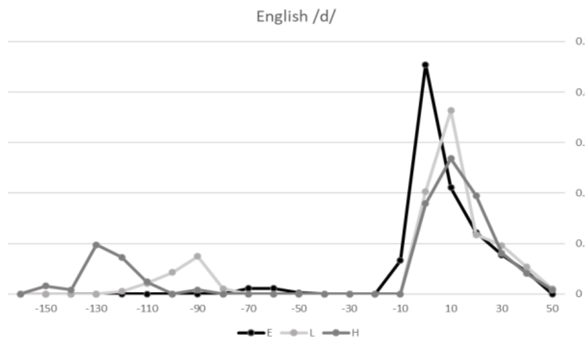


Figure 5. Surfaces of underlying /d/ in English

As for underlying /d/ in English, unaspirated voiceless [t] is the major surface. [d] can phonetically be the surface, so English natives still have little distribution of negative VOTs. Besides, in Mandarin-Min bilinguals' production of /d/, similar patterns with /b/ are found. They reach the negative VOT peaks at －90 and－130 ms, presenting a vast distribution of [b] around －80 ~ －140 ms. It is found that degree of Min fluency significantly influences the negative VOT peaks ($p < 0.05$).

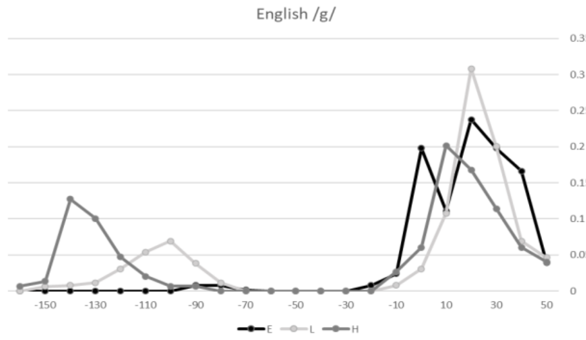


Figure 6. Surfaces of underlying /g/ in English

For the greater jaw movements, VOTs of [k] for underlying /g/ are higher than those of [p] for /b/ and [d] for /t/. Mandarin-Min bilinguals' VOT patterns of underlying /g/ thus present a more separate distribution between two surfaces, unaspirated voiceless stops [k] and voiced stops [g]. As to [g] for /g/, the peaks and the major distributions of negative VOTs by Mandarin-Min bilinguals are higher than [b] and [d]. The [d] distributions of high Min-fluency bilinguals reach the peak at －140ms; in comparison with low Min-fluency bilinguals, their intervals between the burst release and glottal vibration generally take shorter, around －90 ms. Data also reaches statistical significance ($p < 0.05$).

4.3 Crosslinguistic Comparison

In the study, crosslinguistic influences on VOTs are mainly shown in English contexts. Negative VOT patterns in English contexts show a complicated crosslinguistic influence across Mandarin, Min, and English. Different Min fluency levels further distinguish negative VOT patterns. Group L, in which Min is less dominant, shows a shorter negative VOT than Group H. For their L1, Mandarin, has no voiced stops in the inventory, their negative VOTs are not as longer/many as those produced by Group H. In general, low Min-fluency subjects have more linguistic transfers from Mandarin to English contexts.

Moreover, aspirated stops in English contexts also provide informative divergences in VOT values. Mandarin-Min bilinguals with high and low Min fluency present longer VOTs for [p^h, t^h, k^h], since Mandarin and Min contexts, in preference to English, are sensitive to aspiration contrasts. This finding corresponds with Chao and Chen's (2008) observation. Overall, English contexts offer comparative information about crosslinguistic influences on VOTs.

5 Conclusion

The present study conducts a crosslinguistic comparison between languages with voicing and/or aspiration contrasts. Results reveal the linguistic transfer of VOT appears mostly in English contexts. More aspiration, with higher VOT values, has been made by Mandarin-Min bilinguals for aspirated voiceless stops than by English natives, since stop contrasts are well constructed by aspiration in Mandarin and Min. Besides, Mandarin-Min bilinguals present two variations of English voiced stops, phonetically voiced and unaspirated voiceless. Different levels of Min fluency are found to influence speakers' tendencies. Low Min-fluency subjects produce more [p, t. k] for /b, d, g/ and shorter negative VOTs, as their dominant language, Mandarin, originally has no negative VOTs in the phonological inventory and phonetic realization. The findings generally demonstrate a clear crosslinguistic influence on VOT patterns.

Acknowledgments

The study has been supported by a grant from College of Foreign Languages & Literature, National Chengchi University. Besides, the author would like to thank Hsiu-Jung Li and Chih-I Chuang for scoring subjects' Min fluency.

References

Boersma, P. (2006). Praat: doing phonetics by computer. *http://www. praat. org/*.

Chao, K.-Y., & Chen, L.-m. (2008). *A cross-linguistic study of voice onset time in stop consonant productions.* Paper presented at the International Journal of Computational Linguistics & Chinese Language Processing, Volume 13, Number 2, June 2008.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics, 27*(2), 207-229.

Chuang, C.-W. (2021). *Mandarin Speakers' Acquisitions and Representations of Flapping in American English in An ESL Context: A Perception and Production Study.* Paper presented at the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).

Docherty, G. J. (2011). The timing of voicing in British English obstruents. In *The Timing of Voicing in British English Obstruents*: De Gruyter Mouton.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of speech and hearing research, 18*(4), 686-706.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384-422.

Liu, D. (2017). The Acquisition of English Word Stress by Mandarin EFL Learners. *English Language Teaching, 10*(12), 196-201.

Port, R. F., & Rotunno, R. (1979). Relation between voice ‐ onset time and vowel duration. *The Journal of the Acoustical Society of America, 66*(3), 654-662.

Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of phonetics, 7*(2), 197-204.

A Study of Re-generating Sentences Given Similar Sentences that Cover Them on the Level of Form and Meaning

Hsuan-Wei Lo, Yifei Zhou, Rashel Fam, Yves Lepage
Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
hsuanweilo@akane.waseda.jp, yifei.zhou@ruri.waseda.jp,
fam.rashel@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract

In this paper, we define a sentence re-generation task: re-generate a sentence given a set of sentences that cover it. Due to the absence of a dataset to perform this task, we firstly build three language resources of a new type containing more than 4 million annotated sentences. They contain sentences annotated with similar sentences from the same corpus, that cover them on the level of form or meaning. We then perform the sentence re-generation task on the newly produced language resources using two approaches. The first one is a naïve approach where we rely on a language model to reorder the covering parts. The second one is a neural approach where we treat the sentence re-generation task as a translation task from a sequence of covering parts to the respective original sentence. The performance of the systems is evaluated on the level of form and meaning according to the type of covering used to re-generate the sentence. On the level of form, experimental results show that the neural approach outperforms the baseline in edit distance with up to 40% lower scores. However, in BLEU scores, the neural approach is similar or worse than the baseline. On the level of meaning, the neural approach always performs better than the baseline with average scores of 89% BERTScore.

1 Introduction

Translation memories (TM) are used by translators to retrieve close sentences to a given source sentence as hints for translations. It has been shown that similar sentences or informative sentences retrieved from TM can significantly boost the performance of neural machine translation (NMT) systems (Xu et al., 2020; Bulté and Tezcan, 2019a). The idea of a set of sentences covering a given input sentence can be introduced to combine TM with machine translation. (Liu and Lepage, 2021) proposed a twofold-objective approach to retrieve similar sentences based on an input sentence.

To assess the quality of sentence coverage, we define a sentence re-generation task:

> Re-generate an input sentence back given a set of sentences that cover it on two levels, form or meaning.

Yet, there is a lack of data containing for this specific task. The needed data should contain sentences similar in form and meaning along with the covering parts annotated for each input sentence from some corpus. Therefore, we build and release such language resources to meet the requirement of this new task. We then test a naïve baseline and a neural approach on the newly built resources to perform the sentence re-generation task. The two above aspects are the contributions of our work.

The paper is organized as follows. Section 2 explains the notion of coverage of a sentence by other sentences. Section 3 describes the new language resources created. Section 4 explains the two approaches to the sentence re-generation task and the experiment protocol. Section 5 presents the experimental results and provides an analysis. Section 6 gives the conclusion and possible future directions.

Input	*two young toddlers outside on the grass.*
Form	*two white bunnies are* ***outside on the grass****.*
	two young *girls playing outside on the playground.*
	two ***toddlers*** *posing for the camera.*
Meaning	*there is a toddler playing* ***on a*** *playground.*
	two kids are laughing ***in the grass****.*
	a baby is sitting ***outside in grass****.*
	a man ***and a young toddler are playing outside in the grass****.*

Figure 1: An input sentence (Input) and the list of sentences that cover it in form (Form) and meaning (Meaning). All sentences are from the English part of Multi30K.

2 Coverage of a Sentence in Form and Meaning

To produce the newly built released language resources[1], we use a tool for the retrieval of sentences similar to an input sentence that has the claimed twofold objective of

- maximising the coverage of the input sentence in both form and meaning,
- while minimising the number of retrieved sentences.

The tool is implemented as a Python package (Liu and Lepage, 2021). Basically, the input is a sentence (the input sentence) and a corpus, and the output is two lists of sentences extracted from the corpus which are similar to the input sentence. The two output lists are:

- a list of sentences similar in form to the input sentence that cover it as much as possible, and
- a list of sentences similar in meaning to the input sentence that cover it as much as possible.

The lengths of these two lists are not necessarily equal.

2.1 Coverage of a Sentence in Form

By covering an input sentence in form, we mean that the input sentence is covered by sequences of words found in the sentences from the list of sentences output by the retrieval process. The tool ensures coverage by recursive sub-string matching. A sentence with the longest possible sub-string in common with the input sentence is first retrieved. Then, the remaining parts of the input sentence are explored recursively by the same retrieval procedure.

The sentences marked Form in Figure 1 illustrate this. The input sentence is the English phrase *two young toddlers outside on the grass.* The corpus is the English part of Multi30K. The longest sub-string in common with another sentence in the corpus is ***outside on the grass***, found in the sentence *two white bunnies are outside on the grass.* The process then further retrieves sentences that have sub-strings in common with the remaining parts of the input sentence. Sentences with the longest possible sub-strings in common are first retrieved. This explains why the next sentence contains a sub-string of two words in common with the input sentence (***two young***), while the third one has only a sub-string of one word in common (***toddlers***).

An almost total coverage of the input sentence is obtained by combining the sub-strings. Indeed the coverage is not total, as the full stop is not covered. All together, but in a different order, the three sub-strings ***outside on the grass***, ***two young***, and ***toddlers*** make back the entire input sentence *two young toddlers outside on the grass* (again without the full stop).

2.2 Coverage of a Sentence in Meaning

By covering a sentence in meaning, we mean finding portions of the input sentences that are semantically close to sentences from the corpus. Semantic closeness or similarity is achieved through vector cosine similarity of sub-word, word, or word sequence embeddings.

The four sentences marked Meaning in Figure 1

[1] `http://lepage-lab.ips.waseda.ac.jp` → Projects → Kakenhi Kiban C 18k11447 → Experimental data

are examples of covering sentences for the same input sentence as in the previous section. The sub-strings ***on a***, ***in the grass***, and ***outside in grass*** cover *on the grass* in the input sentence. Notice that semantic similarity does not necessarily mean synonymy.

3 Data

We use three corpora that contain diverse topics that may lead to different performance. For each sentence in the three corpora, we retrieve covering sentences from the same corpus. This is performed, of course, after removing the input sentence, so as to prevent the input sentence from being retrieved. The sentence coverage in the newly produced language resources is shown by annotating the parts in common, in form or in meaning, using an XML-like tags format.

3.1 Original Corpora Used to Produce the Released Language Resources

We use the following three corpora to produce three language resources:

- **Multi30K**[2] (Elliott et al., 2016; Elliott et al., 2017; Barrault et al., 2018) is a collection of image descriptions (captions). This corpus is available in four languages: Czech, English, French and German. In this work, we only use the English part of the corpus. This corpus is heavily used in multilingual image description and multimodal machine translation tasks.

- **Tatoeba**[3] is a collection of sentences in over 100 languages. The number of sentences per language ranges from 10 to over 100,000 sentences per language. In this work, we only consider the English part of the corpus.

- **ACL ARC**[4] (Bird et al., 2008) is a corpus that contains academic articles in English, in the field of computational linguistics. It is a collection of published papers in conferences associated with or organised by the Association for Computational Linguistics (ACL).

Table 1 shows the statistics of the above corpora. The ACL ARC corpus contains around 2.5 million sentences, while the number of sentences that we use from Tatoeba is a little bit more than 1.5 million. Multi30K has a relatively small number of sentences in comparison to ACL ARC and Tatoeba: 30,000, as the name says. The vocabulary of ACL ARC is 100 times larger than that of Multi30K. Tatoeba has the shortest average length of sentences in comparison with the other two corpora. ACL ARC also has around two times longer sentences in both, words and characters, compared with Multi30K.

Looking deeper into the use of words, we observe the following phenomena. In types as well as in tokens, words in ACL ARC are around two times longer than in Multi30K. ACL ARC has a high ratio of hapaxes with more than 60%, in comparison to Multi30K with just above 40% and Tatoeba with a little bit more than 45%. These phenomena are mostly caused by the characteristics of the sentences contained in the corpora themselves. Academic articles are more likely to contain longer and more specialized terms. In contrast, in Multi30K, which contains image captions, sentences are shorter and contain more frequent words, which are in trend, shorter than scientific terms.

3.2 Format of the Released Language Resources

The format of the released language resources follows standard practice in Natural Language Processing, where language pieces appear as raw data, separated by tabulations and with annotations by XML-like tags. Each resource consists of a unique file containing a sequence of sentences on each line. On each line, the input sentence comes first, then the list of sentences for coverage in form, and finally the list of sentences for coverage in meaning. Each sentence is separated from the next one by a tabulation.

The sub-strings or parts in common with the input sentence in the retrieved sentences are identified by XML-like tags. There is only one tag used: `coverage`. It takes one attribute `type` with two possible values: `formal` and `semantic`. For instance, the first sentence in the Form part of Figure 1 appears in the language resource produced from the English part of the Multi30K corpus as follows:

[2] `https://github.com/multi30k/dataset`

[3] `https://tatoeba.org`

[4] `https://catalog.ldc.upenn.edu/docs/LDC2009T29/lrec_08/`

	Multi30K	Tatoeba	ACL ARC
# of sentences	30,014	1,519,509	2,491,483
# of tokens	390,843	11,561,489	57,585,605
# of types	10,376	160,454	1,083,298
Avg. token length	3.87±2.40	4.26±2.19	5.34±3.14
Avg. type length	6.93±2.41	8.06±2.74	8.76±4.28
Type-Token-Ratio	0.03	0.01	0.02
Hapax ratio (%)	41.94	47.24	62.55
# of char./sent.	62.38±20.37	39.03±23.52	145.48±73.26
# of words/sent.	13.19±4.17	9.55±4.97	27.45±14.06

Table 1: Statistics on the original data (the average values are given with standard deviation after the ± sign).

	per input sentence on average	Multi30K	Tatoeba	ACL ARC
Form	# of retrieved sent.	5.43±2.21	3.29±2.32	5.79±3.77
	# of char./sent.	61.9±20.54	44.39±37.33	160.99±78.29
	# of words/sent.	13.51±4.18	10.56±7.54	29.96±14.67
	# of char. in coverage	17.97±13.33	16.19±15.65	35.50±46.24
	# of words in coverage	3.91±2.91	3.87±3.49	6.71±8.75
	Individual coverage (%)	32.08	44.71	27.42
	Cumulative coverage (%)	87.50	85.48	63.14
Meaning	# of retrieved sent.	2.05±1.21	1.52±0.82	2.70±1.42
	# of char./sent.	59.18±19.82	36.90±26.27	148.29±76.96
	# of words/sent.	12.84±4.09	9.18±5.38	27.60±14.44
	# of char. in coverage	33.57±23.49	20.14±16.87	95.19±92.52
	# of words in coverage	7.24±4.77	4.51±3.60	18.19±17.04
	Individual coverage (%)	86.58	82.43	82.93
	Cumulative coverage (%)	89.42	89.52	84.91

Table 2: Statistics on the data produced (the average values are given with standard deviation after the ± sign).

```
two white bunnies are
<coverage type=
"formal">outside on the
grass.</coverage>
```

Although the order of retrieved sentences in the released language resources is fixed as mentioned above, the order is theoretically free, because the values of the attribute `type` give the type of coverage.

3.3 Ratio of Coverage

The coverage of a sentence is the length of the sequence of words that is similar to the input sentence. We find that the average length of semantic coverage is higher than that for formal coverage. This is indicated in Table 2 by the rows # of char. in coverage and # of words in coverage. For the Tatoeba corpus, the average length of semantic coverage is 1.2 times that of formal coverage, while this ratio is 2 for Multi30K and roughly 3 for ACL ARC.

3.3.1 Ratio of Coverage in Form

We measure the ratio of coverage in form by counting the number of identical word sequences in the retrieved sentences covering the input sentence. There are two kinds of ratios of coverage per input sentence: individual coverage and cumulative coverage. The individual coverage computes the length of coverage of a retrieved sentence against the length of the input sentence. We report the average individual coverage over all retrieved sentences. In contrast to the average individual coverage, cumulative coverage is the ratio of the sum of the lengths of all substrings in the retrieved sentences over the length of the input sentence. Table 3 illustrates these ratios on several example sentences.

Table 2 shows that the average individual coverage in form is higher for Tatoeba than Multi30K and ACL ARC. This is not surprising as Tatoeba is known to be repetitive. In terms of cumulative coverage, an input sentence can be almost 85% covered by the retrieved sentences in Multi30K and Tatoeba, whereas it is only covered by 63% in ACL ARC.

3.3.2 Ratio of Coverage in Meaning

To compute the similarity between the input sentence and its retrieved similar sentences in meaning, we use the F1 value of BERTScore[5] (Zhang et al., 2020). The individual coverage in meaning is defined as the BERTScore of the semantic coverage per retrieved sentence with the input sentence. Furthermore, we use the concatenation of the sequences of coverage in all retrieved sentences to calculate the cumulative coverage.

Table 2 shows that the ratio of coverage is up to 80% in all three corpora, Multi30K, Tatoeba, and ACL ARC. Some example results for the calculation of coverage ratio in meaning are shown in Table 3.

4 Experiments

We carry out experiments on sentence re-generation with the newly produced language resources introduced in Section 3. For each resource, we divide the dataset into training and test sets, 90% and 10% respectively. Thus, we have a training set and a test set for each Multi30K, Tatoeba, and ACL ARC.

To illustrate the task, let us look at the example in Figure 1. The task consists in re-generating the sentence marked Input, from the only given of the three sentences marked Form or the four sentences marked Meaning. The covering parts are shown in boldface and are indicated by tags in the raw dataset.

We consider two approaches to address this task. The first one is a naïve approach which uses a language model to perform reordering of the covering parts. This is our baseline. The second one is the neural approach where we treat the sentence regeneration task as a translation task from the covering parts into the original sentence. The performance is evaluated separately according to the type of coverage used. Therefore, we have a distinct evaluation on each of the levels of form and meaning.

4.1 Baseline

As said above, the baseline, which is a naïve approach, just reorders the covering parts. It first extracts the covering parts in the retrieved sentences based on the tags (see Section 3.2). The covering parts are ordered in all possible permutations. From this set of all possible permutations, we select the output as being the permutation with the lowest perplexity according to a language model.

[5] `https://github.com/Tiiiger/bert_score`

Multi30K		Coverage (%)		
		indiv	avg	cum
Input	*two young toddlers outside on the grass.*			
Form	*two white bunnies are* ***outside on the grass****.*	50.00	29.17	87.50
	two young *girls playing outside on the playground.*	25.00		
	two ***toddlers*** *posing for the camera.*	12.50		
Meaning	*there is a toddler playing* ***on a*** *playground.*	82.97	90.12	90.32
	two kids are laughing ***in the grass****.*	92.51		
	a baby is sitting ***outside in grass****.*	90.89		
	a man ***and a young toddler are playing outside in the grass****.*	94.10		

Table 3: Example results for coverage ratio on the three corpora. The individual coverage (indiv) is for each individual sentence. The average coverage (avg) is the arithmetic mean over all individual coverage scores. The cumulative coverage (cum) is the coverage of all sub-strings in the retrieved sentences over the input sentence. Its maximal value is 100%.

Corpus	Approach	Accuracy (%)	Edit distance		# of chars per sent.	# of words per sent.	BLEU points
			in chars	in words			
Multi30K	Naïve	0.30	51.76	12.05	99.28	20.76	**43.60**
	Neural	**7.90**	**29.38**	**6.76**	55.06	12.01	39.61
Tatoeba	Naïve	0.44	33.83	8.37	58.47	12.80	45.05
	Neural	**24.64**	**19.70**	**4.65**	34.74	8.61	**45.57**
ACL ARC	Naïve	0.00	161.31	35.19	212.32	38.83	**23.53**
	Neural	**0.27**	**103.86**	**23.30**	103.47	20.34	8.76

Table 4: Evaluation on the level of form.

Corpus	Approach	# of chars per sent.	# of words per sent.	BERTScore (F1)
Multi30K	Naïve	69.41	14.82	0.86
	Neural	43.43	10.19	**0.91**
Tatoeba	Naïve	30.93	6.69	0.85
	Neural	28.14	7.95	**0.90**
ACL ARC	Naïve	287.11	54.98	0.84
	Neural	91.71	20.70	**0.86**

Table 5: Evaluation on the level of meaning.

4.1.1 Permutation of Covering Parts

We extract the covering parts from the retrieved sentences. These covering parts are to be found between the tags mentioned in Section 3.2. As an example, let us suppose that we have five sentences in the sentence coverage on the level of form for one input sentence. This gives five sub-strings between tags that we extract, which are the covering parts.

Carrying on with the situation of five retrieved sentences, we permute the five covering parts in all possible orders. This gives us 120 possible permutations, i.e., $n!$, where n is the number of covering parts, 5.

4.1.2 Selection by Language Model

We apply a language model to score all possible permutations. We use kenLM[6] (Heafield et al., 2013) for that, and use modified Kneser-Ney smoothing without pruning, for smoothing. In our experiments, we train the language model on the training set. We thus built 3 language models, one on each of the three different corpora: Multi30K, Tatoeba, and ACL ARC.

In our previous example of 120 combinations, applying the language model to each of the 120 combinations delivers a score for each of the combinations. We select the best combination, that is the one with the lowest score among the 120 combinations. Some examples of results are given in Table 6 for each of the three corpora.

4.2 Neural Approach

In a second, more modern and less naïve approach, we treat the sentence re-generation task as a translation task where:

- the source channel is the covering parts contained in the sentence coverage, and
- the target channel is the original sentence which we would like to re-generate.

Similar to the baseline, we only consider the covering parts as the input for the neural approach. Thus, we first extract the covering parts from the sentence coverage according to the tags (see Section 3.2). We then train the Transformer model on the training set.

4.2.1 Preprocessing Dataset

As mentioned above, we extract the covering parts from the retrieved sentences. The covering parts are concatenated as an input sequence to the neural approach. For example, in Figure 1, we use the covering parts on the level of form, "*outside on the grass*", "*two young*" and "*toddlers*" as an input sequence for the input sentence "*two young toddlers outside on the grass*". The source and target channels of the neural approach are shown as follows:

- source channel: "*outside on the grass two young toddlers*"
- target channel: "*two young toddlers outside on the grass.*"

All sentences are tokenized using SentencePiece[7] to break the sentences into sub-words. The subword model is known to improve the performance of the natural language generation systems (Kudo and Richardson, 2018). This tool is an unsupervised text tokenizer (encoding) for neural networks, especially in the text generation system.

4.2.2 Transformer: Open-NMT

We train a Transformer model provided in OpenNMT-py[8] (Klein et al., 2017) on each of the three language resources mentioned above to perform the sentence re-generation task. We keep the test set as it is (10%) and divide the original training set (90%) into 80% as a training set and 10% as a validation set to train our Transformer model. Next, we select the best-trained model to perform the sentence regeneration task in terms of the perplexity score given by the transformer model on the validation set.

5 Results and Analysis

We evaluate the performance of the baseline and the neural approach on the test sets on each of the three corpora. As mentioned in Section 4, the evaluation is performed on two levels, form and meaning, based on the type of coverage used. On the level of form, we use accuracy, edit distance, and BLEU scores as evaluation metrics. To measure the performance on the level of meaning, we use BERTScore. The overall results are given in Table 4 and Table 5.

[6] `https://github.com/kpu/kenlm`

[7] https://github.com/google/sentencepiece

[8] https://github.com/OpenNMT/OpenNMT-py

Corpus	Level	Approach	Sentence	BLEU	BERT-F1
Multi30K	-	-	*two young toddlers outside on the grass .*	-	
	Form	Naïve	*two young toddlers outside on the grass*	86.69	-
		Neural	*two young toddlers outside on the grass .*	100.00	-
	Meaning	Naïve	*on a and a young toddler are playing outside in the grass in the grass outside in grass*	-	0.91
		Neural	*a toddler is playing with a toy in the grass .*	-	0.93
Tatoeba	-	-	*I'd be unhappy, but I wouldn't kill myself.*	-	
	Form	Naïve	*but I would n't be unhappy, wouldn't kill I'd be unhappy, but I*	36.41	-
		Neural	*I'd be unhappy, but I wouldn't kill.*	79.56	-
	Meaning	Naïve	*but I don't intend to kill myself*	-	0.86
		Neural	*I don't intend to kill myself, but I don't want to kill myself.*	-	0.92
ACL ARC	-	-	*Single word may have different meanings under different situations.*	-	
	Form	Naïve	*under different situations Single word*	18.39	-
		Neural	*Sometimes different situations under different situations.*	28.32	-
	Meaning	Naïve	*have different meanings . g . word followed by comma) can also be addressed through truecasing .*	-	0.85
		Neural	*In other words, words may have different meanings.*	-	0.92

Table 6: Examples of sentences re-generated by the naïve baseline and neural approaches on the three corpora using coverage on the level of form and meaning. A grey background indicates the original sentence from the corpus, a white background is for the re-generated sentence.

5.1 Evaluation on the Level of Form

The first metric used to evaluate the performance of the system on the level of form is accuracy. Accuracy is defined as the proportion of exact match between the reference and the re-generated sentence (just a full stop missing counts as zero). The naïve baseline's performance is close to zero in terms of accuracy on all of the corpora: 0.30% on Multi30K, 0.44% on Tatoeba, and 0.00% on ACL ARC. There is no correct sentence re-generated by the baseline on the ACL ARC. However, the neural approach has higher accuracy than the baseline on the three corpora: 7.90% on Multi30K, 26.64% on Tatoeba, and 0.27% on ACL ARC. The difference is pretty high, particularly on Tatoeba.

A finer view is given by the use of the Levenshtein edit distance (Levenshtein, 1966)[9] to perform the formal evaluation. The Levenshtein edit distance involves three different operations: insertion, deletion, and substitution. Each operation counts as one. Table 4 gives the results of the application of the Levenshtein edit distance at two levels of granularity, that of characters and that of words. The edit distance of the baseline is very close to the average length of the original sentence. This means that almost all of the words in the re-generated sentence need to be modified. We also observe that the neural approach achieves around 40% lower edit distance than the baseline. Looking at some examples of the re-generated sentences in Table 6, the Transformer model has removed repeated words or grammatical errors such as punctuation. This makes re-generated sentences closer to the reference sentences in terms of edit distance.

We thus measure the extent to which groups of words might be in the correct order. To this end, we use BLEU (Papineni et al., 2002), in its SacreBLEU[10] (Post, 2018) implementation. BLEU evaluates the similarity between one or several original references, and a candidate sentence. Higher BLEU scores indicate higher similarity, with 100 being the maximum. Table 4 shows that the base-

[9] https://github.com/roy-ht/editdistance

[10] https://github.com/mjpost/sacrebleu

line is able to obtain around 44 BLEU scores for Multi30k and Tatoeba, twice as much as for ACL ARC (23.53). BLEU scores of above 40 reflect the fact that some sequences of words are shared between the re-generated sentence and the reference sentence, i.e., not all words are scrambled in a completely different order. We also notice that the neural approach gets similar (Tatoeba) or lower (Multi30K and ACL ARC) BLEU scores than the baseline. This indicates that the Transformer model missed half of the portion of correct words in the correct position. For ACL ARC, the neural approach seems to change most of the content of the sentence into a completely different sentence which leads to a low BLEU scores (8.76).

5.2 Evaluation on the Level of Meaning

To evaluate the performance on the level of meaning, we use BERTScore (similar to what we did in Section 3.3.2). It computes the cosine similarity between pair-wise tokens in form of the re-generated sentence and the original sentence using a pre-trained BERT embedding model. BERTScore provides precision, recall, and F1, measured by the weighted maximum similarity scores. Here, we only consider the F1 score which is the harmonic mean of recall and precision. A value of 1 for the F1 score means that the meaning of the re-generated sentences and reference sentences are the same. Table 5 shows that F1 scores on Multi30K, Tatoeba, and ACL ARC are in the range of the mean of 0.87 to 0.92. This shows that the re-generated sentences are 85% semantically close to the original input sentences, according to BERTScore. Overall, the neural approach performs better than the baseline with an average score of 0.89. This does not necessarily show that our re-generated sentences are close to the reference sentence, as shown in Table 6.

5.3 Discussion

Our experimental results show a configuration where scores in accuracy close to zero and large Levenshtein edit distances are seemingly in contradiction with the reasonably high scores in BLEU and the excellent scores in BERTScore. Such an experimental configuration asks the question of what the adequate metrics are to reflect the fact that, although almost all the expected words are present, although some sequences of words are correct and match exactly the input sentence, as a whole, the re-generated candidate sentences produced by our naïve method are far away from the input sentences.

6 Conclusion

We defined a sentence re-generation task: regenerate an input sentence back given sentences that cover it on two levels, form and meaning. For this task, we built a new type of language resource produced from three different corpora: Multi30K, Tatoeba, and ACL ARC. Altogether, this represents over 4 million sentences annotated with similar sentences that cover them, and in which the covering parts are tagged. We released three resources.

We carried out experiments on this new type of resource using two approaches: a naïve approach of sequence reordering using a language model, and a neural approach that treats the sentence regeneration task as a translation task from covering parts to the original sentence. The experiments were performed on both the level of form and meaning. The performance of the systems was evaluated according to the type of coverage used. On the level of form, experimental results showed that the neural approach performed up to 40% better in terms of accuracy and edit distance. However, it performed similarly or lower in terms of BLEU. On the level of meaning, the neural approach achieved a higher BERTScore than the baseline by a margin of 4%.

For future work, as for the released language resources, their quality can be improved. A higher individual coverage percentage (mentioned in Section 3) is needed, particularly for the ACL ARC corpus, so that original sentences are enough covered. In addition, the calculation of coverage percentage on meaning needs to be revised since the average of 83% of individual coverage percentage was not reflected well in our evaluation on the level of meaning.

We also consider releasing similar resources for other languages than English. We believe that this new type of resource can be used for various types of NLP tasks such as language reference (Vossen et al., 2020), text generation (Nan et al., 2021), or the integration of translation memories with machine translation (Bulte and Tezcan, 2019b).

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of LREC 2008*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Bram Bulté and Arda Tezcan. 2019a. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of ACL 2019*, pages 1800–1809.

Bram Bulte and Arda Tezcan. 2019b. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of ACL 2019*, pages 1800–1809, Florence, Italy, July. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL 2013*, pages 690–696, Sofia, Bulgaria, August. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP 2018: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-doklady*, 10(8):707–710, February.

Yuan Liu and Yves Lepage. 2021. Covering a sentence in form and meaning with fewer retrieved sentences. In *Proceedings of PACLIC 35*, pages 1–10, November.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of NAACL 2021: Human Language Technologies*, pages 432–447, Online, June. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, July. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Piek Vossen, Filip Ilievski, Marten Postma, Antske Fokkens, Gosse Minnema, and Levi Remijnse. 2020. Large-scale cross-lingual language resources for referencing and framing. In *Proceedings of LREC 2020*, pages 3162–3171, Marseille, France, May. European Language Resources Association.

Jitao Xu, Josep M Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of ACL 2020*, pages 1580–1590.

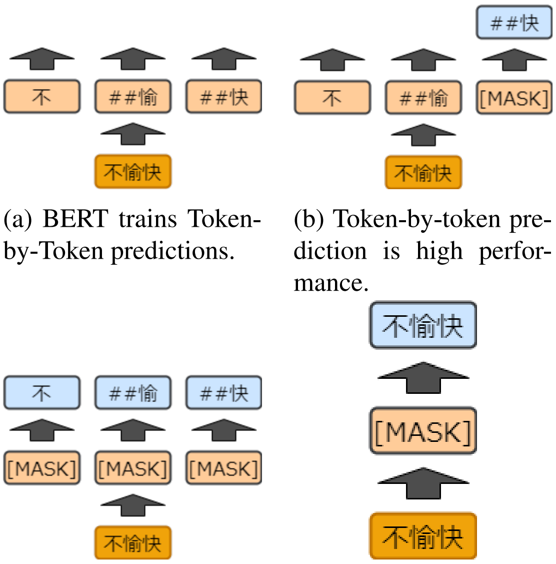
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR 2020*.

Vocabulary expansion of compound words for domain adaptation of BERT

Hirotaka Tanaka **Hiroyuki Shinnou**
Department of Computer and Information Sciences Ibaraki University
Hitachi, JAPAN
`{ 22nd304a, hiroyuki.shinnou.0828 } @vc.ibaraki.ac.jp`

Abstract

Pretraining models such as BERT, have achieved high accuracy in various natural language processing tasks by pretraining on a large corpus and fine-tuning on downstream task data. However, BERT trains token-level inferences, which make it difficult to train unknown or compound words that are split by byte-pair encoding. In this paper, we propose an effective method for constructing word representations in vocabulary expansions for such compound words. The proposed method assumes domain adaptation by additional pretraining and expands the vocabulary by embedding a synonym as an approximate embedding of additional words. We conducted experiments using each vocabulary expansion method and evaluated these experiments for their accuracies in predicting additional vocabularies in the masked language model.



(a) BERT trains Token-by-Token predictions.

(b) Token-by-token prediction is high performance.

(c) Difficult to predict the entire splitted representation.

(d) The entire representation as a single Token is easily predicted.

Figure 1: Examples of vocabulary expansion targets

1 Introduction

Pre-learning models have significantly improved the performances of various natural language processing systems (Peters et al., 2018)(Radford et al., 2018). Bidirectional encoder representations from transformers (BERT)(Devlin et al., 2019) is pretrained model that consists of a stacked multi-head attention in the Transformer(Vaswani et al., 2017). A BERT outputs word representations that embed the context of the input word sequence.

Pre-learning models have domain adaptation problems, and they perform various downstream tasks with high accuracy by applying models pretrained on a large corpus to the downstream task data. Therefore, if the domain of the downstream task differs significantly from the domain of the pretrained corpus, the solution of the downstream task will have low accuracy. Gururangan et al.(2020) proposed a method for domain adaptation by additional pretraining on a corpus of downstream task domains.

Domain adaptation problems also appear in the vocabulary. The vocabulary covered by the pretrained model depends on the pretrained corpus. Therefore, it must adapt to the vocabulary appearing in the downstream task data through vocabulary expansion. The adapt-and-distill approach by

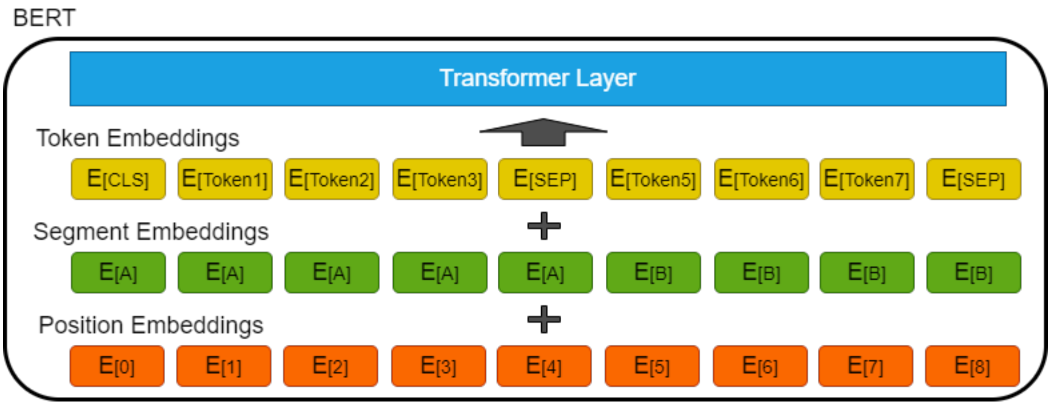


Figure 2: Construction of BERT input embeddings

Yao et al.(2021) and AVocaDo strategy by Hong et al.(2021) are vocabulary expansion methods for domain adaptation in the pretraining models.

BERT trains token-level expressions, and it has difficulty training the expressions of compound words, named entities, and phrases. The Japanese vocabulary for a standard BERT model is token units, and tokenization is performed by morphological analysis and byte-pair encoding. The division of unknown words into known words makes it possible to cover a large number of words using a small vocabulary. However, the masked language model (MLM) in the pretraining task trains for token by token prediction. Therefore, it is difficult to predict an entire representation constructed using multiple tokens (see Figure 1).

Yanada et al.(2020) proposed the pretraining model LUKE for training the representations of entities constructed from multiple words. LUKE provides entity embeddings in addition to ordinary word embeddings and models and trains the relationship between ordinary tokens and entities by the entity-aware self-attention mechanism. However, LUKE requires expensive pretraining.

In this paper, we propose a method to add a vocabulary to the BERT model. In vocabulary expansion, the focus is on the method for constructing word embeddings in the additional vocabulary. By assuming domain adaptation through the additional pretraining of downstream task data, we expect the model to train word embeddings of the additional vocabulary based on approximate vectors.

2 Related work

2.1 Token embeddings for BERT

Tokens obtained from the input sentences are converted to token embeddings. The BERT input vector consists of three embeddings (see Figure 2). Token embeddings represent words. Segment embeddings embed information that identifies each sentence from among the multiple input sentences, and position embeddings represent the token's position in the input.

The final output of BERT is a contextualized word representation. However, the embedding used as input to BERT is unique for each token.

2.2 MLM

MLM is a pretraining method of BERT. The task was to predict the tokens replaced by MASK tokens. The standard approach is to replace 15% of the input tokens. These 15% input tokens include the following replacements:

- 80% are replaced with the special token `[MASK]`.
- 10% are replaced by other random tokens.
- The remaining 10% tokens are kept intact.

3 Vocabulary expansion for BERT

We expand the vocabulary of the pretrained BERT model by adding word embeddings to token embeddings. Therefore, the challenge for the vocabulary expansion methods is to obtain additional word embeddings corresponding to the BERT embedding space.

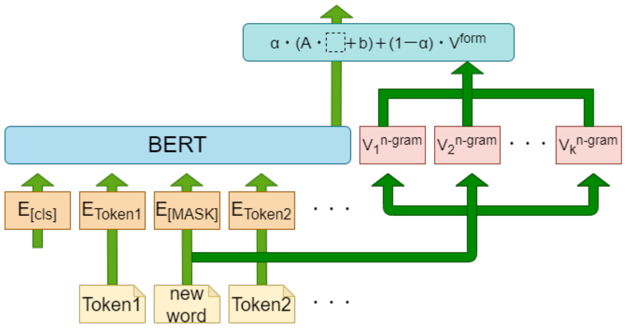


Figure 3: SHALLOW method of BERTRAM

3.1 Static word representations

A method for obtaining additional word embeddings is to use static word representations from distributed representation models such as Word2Vec or fastText(Bojanowski et al., 2017)(Joulin et al., 2016). When a distributed representation model trained on the target word exists, vocabulary expansion is achieved by adding the model to the BERT embeddings. Even when a distributed representation model with trained target words is absent, the calculation cost is lower to train a distributed representation model that includes the target words than to pretrain a BERT model from scratch.

In particular, we first prepare a distributed representation model trained on the additional vocabulary. Next, the transformation model trains a mapping from the distributed representation model to the BERT word embedding based on the vocabulary set commonly trained by them. Mikolov et al.(2013) proposed a method for training the mapping by applying stochastic gradient descent to reduce the mean squared error between the source and target word vectors.

3.2 Mean vector of subwords

The mean vectors of subwords are used to obtain embeddings using only pretrained BERT models. This method is also used in adapt-and-distill approach(Yao et al., 2021).

The target words of the expansion are processed with multiple tokens in BERT. The mean vector of these token embeddings is calculated from the pretrained BERT model. Let the mean vector be the additional word embeddings.

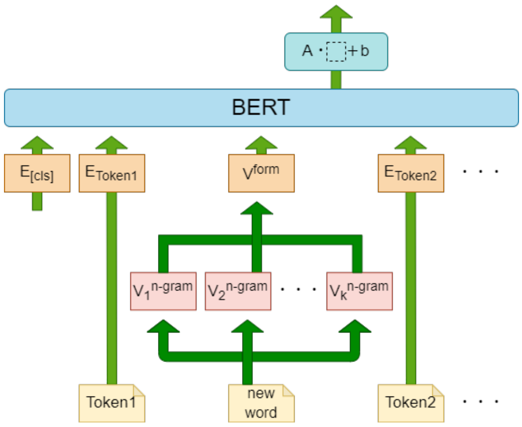


Figure 4: REPLACE method of BERTRAM

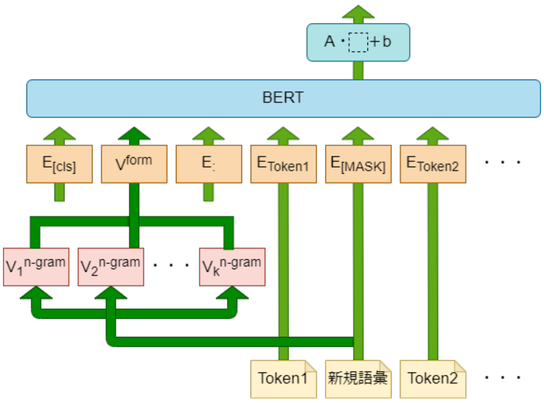
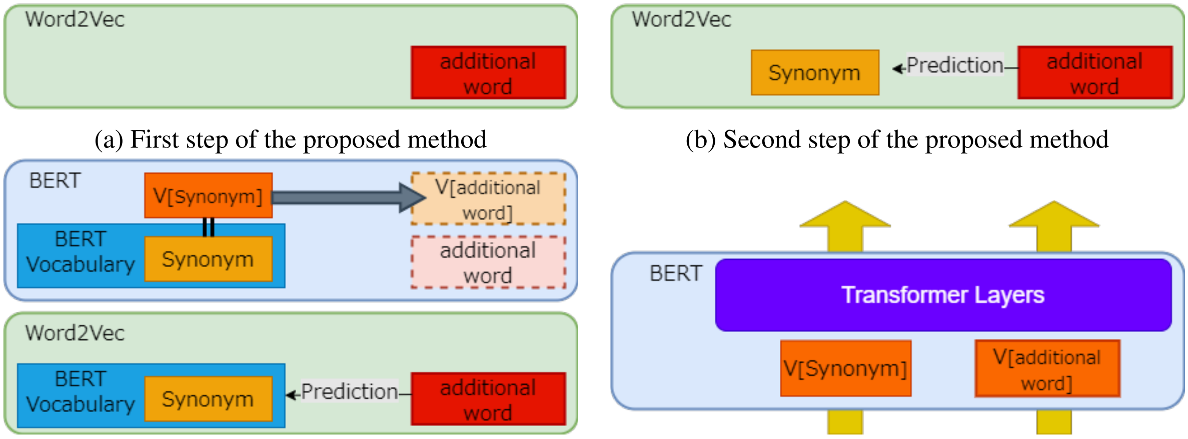


Figure 5: ADD method of BERTRAM

3.3 BERTRAM

Schick and Schütze (2020) proposed BERT for attentive mimicking (BERTRAM) as a method for obtaining additional word embeddings from the output of the pretrained BERT. BERTRAM focuses on adding low-frequency words in the pretrained corpora and trains additional word embeddings in a form-context model. The form model trains on a character basis, whereas the context model trains on a context basis. The form model trains character n-gram embeddings and constructs additional word embeddings from them. The context model trains additional word embeddings by predicting masked additional words in a sentence, which is similar to the method used by the standard BERT model to train contextualized word embeddings.

For the context model construction, Schick et al. tried three methods: SHALLOW, REPLACE, and ADD.



(a) First step of the proposed method

(b) Second step of the proposed method

(c) Third step of the proposed method

(d) Final step of the proposed method

Figure 6: Schematic figure of the proposed method

In the SHALLOW method, the output of the pre-trained BERT for the mask token is obtained as the contextualized word embedding of the target word (see Figure 3).

In the REPLACE method, the embedding by the form model is input to the pre-trained BERT as a target word embedding. The output of the BERT is then obtained as a contextualized word embedding representation of the target word (see Figure 4).

The ADD method is a combination of the SHALLOW and REPLACE methods. First, the target words are replaced with mask tokens. Then, a word embedding using the form model and the word ":" are added at the beginning of the sentence. The resulting sentence is input to the pre-trained BERT, and the embedding corresponding to the mask token in the output is obtained as a contextualized word embedding of the target word (see Figure 5).

In the case that the context model method is SHALLOW, the embeddings of the form-context model are computed as:

$$v_{(w,C)} = \alpha \cdot (A \cdot v^{context}_{(w,C)} + b) + (1-\alpha) \cdot v^{form}_{(w,C)} \quad (1)$$

where w is a target word, C is a sentence set, and $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^{d}$ are trainable parameters.

In the case that the context model method is REPLACE or ADD, the embedding of the form model output is an input to the context model. Therefore, the embedding of the form-context model is obtained by linear transformation of the context model output:

$$v_{(w,C)} = (A \cdot v^{context}_{(w,C)} + b). \quad (2)$$

The loss on training is computed as follows:

$$\| e_w - v_{(w,C)} \| \quad (3)$$

where e_w is the BERT token embedding corresponding to word w.

BERTRAM training is a three-step process. When the context model method is ADD, the training method involves the following steps:

1. Train only the context model using the SHALLOW method.

2. Train only the form model.

3. Train the entire parameters of the BERTRAM model constructed by the context model using the ADD method.

When training BERTRAM with the ADD and REPLACE methods, the SHALLOW method is also used to trian the context model in the first step.

The training parameters of the pretrained BERT model are frozen during all training steps.

In the BERTRAM experiment, the additional vocabulary "____" is added as a special token denoted by "<BERTRAM:____>." When using word embeddings added by BERTRAM, the additional vocabulary in the data is replaced with "<BERTRAM:____>" representations.

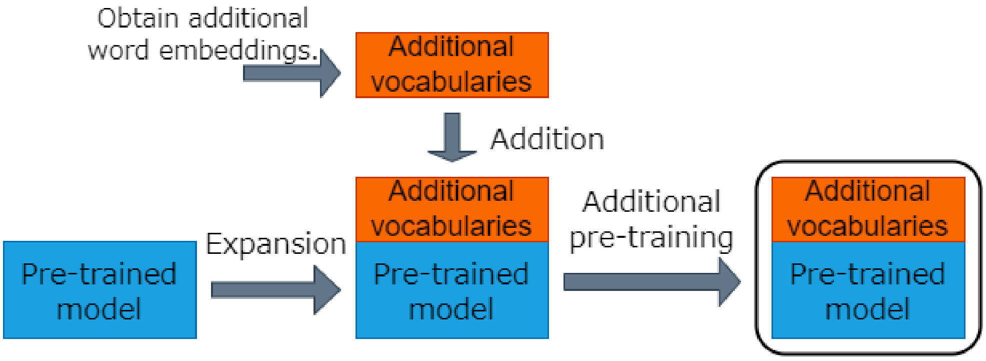


Figure 7: Vocabulary Expansion Process

4 Proposed method

Domain adaptation to the downstream task domain enhances the performance of the prelearning model for downstream tasks. We expect that when adapting the model to the domain using the method of Gururangan et al.(2020) even though the additional word embeddings are approximate, the additional pretraining can learn the appropriate embeddings.

Approximate embeddings of additional words can be obtained from their synonyms. Additionally, the synonyms included in the pretrained BERT vocabulary have their embedded representations already learned by BERT. Therefore, in the proposed method, we add the synonym embeddings included in the pretrained BERT vocabulary to the model as the additional word embeddings. Then, the model additionally pretrains on the downstream task data.

We applied distributed representation models such as Word2Vec, for synonym estimation. The distributed representation model is a model trained on both the additional vocabulary and BERT vocabulary.

At the vocabulary expansion stage, the synonym embeddings and additional word embeddings are identical. By additionally pretraining the entire training parameters, including BERT's token embeddings, the model trains the optimal embedding for each model. Additional pretraining uses the downstream task domain as the training data, which fine-tuning the entire training parameters of the model with the MLM.

In the proposed method (see Figure 6), the model fine-tunes the word embeddings with MLMs based on the synonym embeddings of the additional vocabulary.

The method involves the following steps:

1. A distributed representation model such as Word2Vec, is trained with the additional vocabulary.

2. The distributed representation model estimates the synonyms for the additional vocabulary.

3. The token embeddings of the estimated synonyms in the BERT vocabulary are included as additional word embeddings.

4. The BERT with the expanded vocabulary is additionally pretrained with the MLM on a corpus containing the additional vocabulary.

5 Experiments

5.1 Methodology

In this experiment, we extended the pretrained Japanese BERT vocabulary(see Figure 7). In addition to the proposed method, we compared the method using static word embedding, the method of mean vectors of subwords, and BERTRAM.

The proposed method applies cosine similarity to the similarity between additional vocabularies and synonyms.

The experiments for all methods were conducted under the following conditions:

- MeCab was applied for the morphological analysis.

- The models added to the vocabulary by each method were additionally pretrained on the downstream task domain data by the MLM.

Target words	Synonyms	Similarities
殺人事件	殺人	0.8607
社会科学	人文	0.8482
推理小説	ミステリ	0.8147
自分自身	自分	0.8105
彼女自身	彼女	0.8102
日本文学	国文学	0.8080
新聞記者	記者	0.8036
長編小説	小説	0.8006
短編小説	小説	0.7950
携帯電話	携帯	0.7866
近親相姦	レズ	0.7858
少年時代	幼少	0.7831
金融機関	銀行	0.7758
統合失調症	うつ病	0.7750
精神医学	臨床	0.7687
日常生活	日常	0.7675
成果主義	デフレ	0.5805
地球温暖化	エコ	0.5796
練習問題	文法	0.5764
人生経験	悩み	0.5706
自己満足	信念	0.5691
宣伝文句	キャッチフレーズ	0.5649
参考文献	文献	0.5636
古今東西	落語	0.5500
携帯小説	妄想	0.5477
行政書士	弁護士	0.5416
設定資料集	プラモデル	0.5342
成功法則	生き方	0.5338
日本語版	翻訳	0.5316
自己責任	考え方	0.5269
不完全燃焼	不発	0.5246
裁判員制度	陪審	0.5194

Table 1: Examples of synonyms and similarities

- The evaluation method estimated the new vocabulary by replacing only the new vocabulary with the MASK tokens.
- The evaluation index was the mean reciprocal rank (MRR).
- The evaluation value was the average of five trials with different random seeds.

5.2 Japanese pretrained BERT model

We used the model named cl-tohoku/bert-base-japanese published by the Inui Laboratory at Tohoku University as the pretrained BERT. This model is available in Hugging Face's transformers library. The model was pretrained using the Japanese Wikipedia as the pretraining corpus.

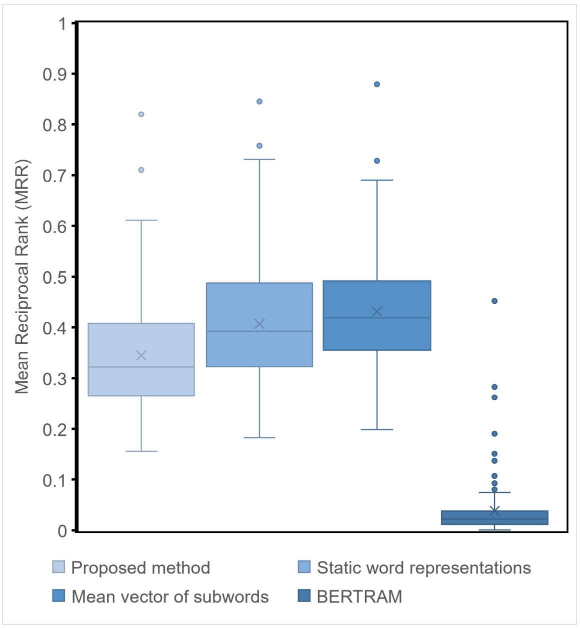


Figure 8: The experimental results

5.3 Distributed representations

For static word embeddings, we used the Japanese Wikipedia entity vectors available at `http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/`. This vector was trained using the Word2Vec model.

5.4 BERTRAM

To learn BERTRAM, we applied the programs adapted to the Japanese language by referring to the author-implemented programs of the BERTRAM paper available at the following two websites: `https://github.com/timoschick/bertram` and `https://github.com/timoschick/form-context-model` . This BERTRAM model was trained on the Japanese Wikipedia.

5.5 Datasets

The dataset for this experiment is sourced from the Amazon Review Corpus. This corpus is available at `https://webis.de/data/webis-cls-10.html`.

This corpus was divided into three domains according to the type of product: Books, DVDs and

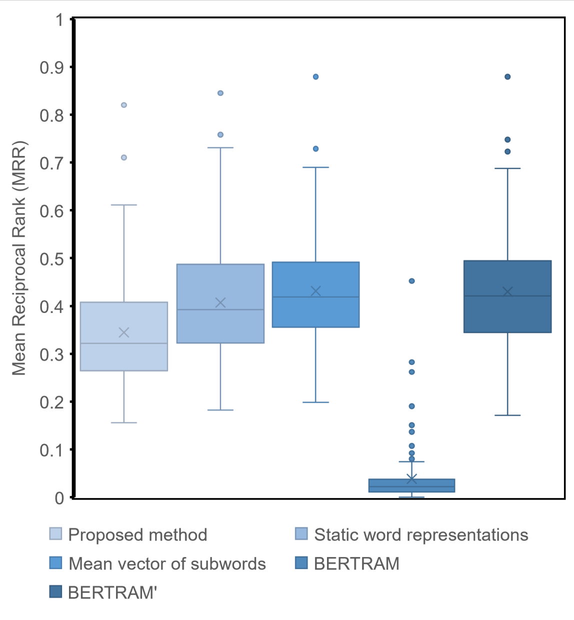


Figure 9: BERTRAM tokenization experiment

Music. In this experiment, we used the Books domain. The corpus included unlabeled reviews and reviews labeled with stars. We used unlabeled data from the Books domain.

We defined the additional target words as having four or more Kanji characters, which were defined as general proper nouns in the Mecab Ipadic NEologd dictionary. There were 149 such words in the Amazon review corpus. These words were divided into multiple tokens in the pretrained BERT and into multiple morphemes in the standard MeCab. The dataset had 100 sentences of the corpus for each word. For each word, 50 sentences were those of the training data and the other 50 were the test data. The training and test data had 7,450 sentences each.

6 Results and discussion

6.1 Experimental results

Examples of synonyms and similarities obtained using the proposed method are shown in Table 1, and this table presents the top 16 and bottom 16 similarities.

The experimental results for each method are shown in Figure 8. This figure presents the MRR in each word as a box-and-whisker plot. Although the proposed method had relatively low accuracy, its prediction accuracy was as good as that of other methods. There was no significant difference between the static word representations and mean vector of the subwords. However, BERTRAM's prediction accuracy was significantly low.

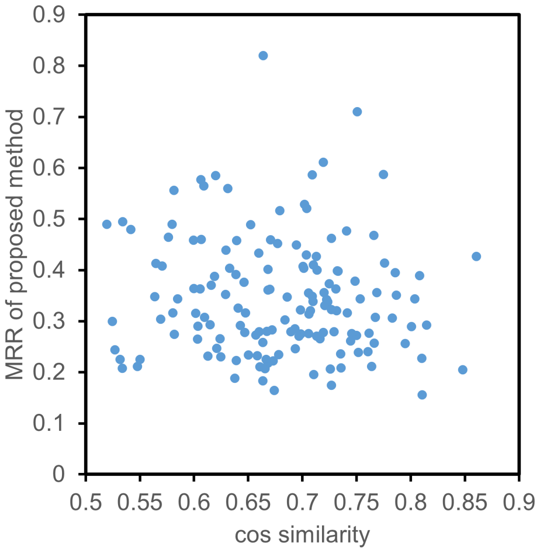


Figure 10: Similarity and accuracy for each word

6.2 Tokenization of BERTRAM

We identify the causes of BERTRAM's experimental results. Tokenization in BERTRAM is different from other methods used for additional vocabulary. BERTRAM initially picks out the additional vocabulary in the sentence as special tokens and splits the sentence before and after it. Then, it tokenizes each of the split sentences.

This tokenization works in English, but it does not work in Japanese. Japanese requires morphological analysis. When BERTRAM's tokenization is applied to Japanese, it morphologically analyzes each segmented sentence. The morphological analyzer does not receive a complete sentence, which results in low analytical accuracy.

We conducted an experiment to confirm this. We let BERTRAM' be the method of morphological analysis of complete sentences similar to other methods. The results of this experiment are shown in Figure 9. BERTRAM' was as accurate as the static word representations and mean vectors of subwords. Therefore, the low accuracy of BERTRAM resulted from tokenization.

6.3 Relationship between similarity and prediction accuracy

To validate the effectiveness of the proposed method for each word, we analyzed the relationship between the similarities and accuracies of the synonyms. A graph that plots the similarities and accuracies for the words is shown in Figure 10. The correlation coefficient between the similarities and accuracies is -0.0815. Therefore, there was no significant correlation between the similarities and accuracies if synonyms.

Using the accuracy of each method for each word, it was not possible to determine the most effective method for all words. Futher analysis is required to determine the specific word features that would be effective for each method.

7 Conclusion

In this paper, we examined how to apply an effective vocabulary expansion method to compound words. We proposed a method using synonyms by assuming domain adaptation with additional pretraining. An analysis of the relationship between the prediction accuracy and similarity of the synonyms to the target word revealed no significant correlation between them. Furthermore, we improved the tokenizer implemented in BERTRAM to apply it to the Japanese language. Our future work will include an analysis of the features of compound words, and we will propose effective methods for compound words. We also propose to extend the method to sequence representations longer than those used in this study.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI.*

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online, July. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 460–470, Online, August. Association for Computational Linguistics.

A comparison of the validity of measurement methods of the general English proficiency by dictation and read-aloud performance

Katsunori Kotani
Kansai Gaidai University
Osaka, Japan

kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi
Ryukokou University
Shiga, Japan

yoshimi@rins.ryukoku.ac.jp

Abstract

This paper compares three classroom-based measurement methods of the general proficiency of English as a second language (GEP): one measures the GEP by dictation performance; another by read-aloud performance; and the other by both dictation performance and read-aloud performance. A learner's GEP has been measured by commercially tests because the reliability and validity of the tests have been well-acknowledged, but the use of the tests has ineffective regarding test-administration cost and test materials. By contrast, measurement of dictation and/or read-aloud needs only a sound file or a text file, and a teacher of English as a second language can choose test materials according to the class contents. This study developed the three GEP-measurement methods using a learner corpus data involving GEP, dictation and read-aloud performance of 50 learners of English as a second language. The experimental results suggested that the proficiency should be measured with both dictation and read-aloud performance.

1 Introduction

In a class of English as a second language (ESL), a teacher should conduct test in a classroom, because classroom testing has benefits to grow students' general English proficiency (GEP). Classroom testing provides a teacher with valuable feedback about students' learning outcome (Roediger et al. 2011). Thus, a teacher can understand how students' GEP grows. In addition, frequent testing encourages students to study (Roediger et al. 2011). Thus, if a teacher provides a test in a classroom, students will increase learning motivation.

A classroom testing method is necessary, because commercially available tests such as Test of English as a Foreign Language (TOEFL) or Test of English for International Communication (TOEIC) have three limitations. First, test fees are expensive for classroom testing, although it can be applicable once in an academic year for confirming learning outcome or using as a placement test to determine a class appropriate for students' proficiency. Second, test administration takes a couple of hours. It is longer than a university class period (90–105 minutes in Japan). Finally, the test material is irrelevant for the ESL classes being taken by students.

A solution to these limitations is to introduce a computer-assisted language testing (Noijons 1994, Suvorov 2013). Here, a learner's GEP was measured by calculating scores for a learner's reading aloud performance and dictation performance. Dictation and read-aloud performances were taken up in this study for four reasons.

First, tests for dictation and read-aloud have been reported to demonstrate the reliability and validity as a GEP test (Kotani & Yoshimi 2018, Kotani & Yoshimi 2020). In addition, dictation and read-aloud tests have also been reported to demonstrate GEP (Irvine, Atai, and Oller, Jr. 1974; Oller 1983; Coniam 1991; Lee 2004; Iino, Yabuta and Thomas 2011; Kazazoglu 2013; Wong and Leeming 2014; Leeming and Wong 2016; Yazdinejad and Zeraatpishe 2019). Given the close relation of GEP with dictation and read-aloud performance, dictation and read-aloud tests are used in Duolingo, commercially available testing for GEP (Wanger 2020).

Second, these tests save an ESL teacher's time and efforts to prepare and administer reading/listening comprehension questions. An ESL teacher only must prepare a sound/text file for dictation/read-aloud. Thus, an ESL teacher can use a dictation/read-aloud test as a quick check for GEP several times in a semester/in every class. In addition, a teacher can choose test materials according to the class contents.

Third, a dictation/read-aloud test is low cost for evaluation. In a dictation test, students' answers can be evaluated by comparing with reference sentences manually by an ESL teacher or automatically by a natural language processing tool to measure edit distance. In read-aloud test, the accuracy of pronunciation can also be evaluated by comparing with reference pronunciation manually by an ESL teacher or automatically by a sound recognition tool (Fu et al. 2020).

Finally, both dictation and read-aloud tasks are also effective from the pedagogical viewpoint. Kojima and Ota (2012) investigated the effect of dictation, read-aloud and shadowing by comparing test results between a pre-test and a post test of a semester. The results indicated that dictation could improve listening ability than shadowing.

Previous research can be classified into two categories. One examined the correlation of GEP with dictation or read-aloud performances (Irvine, Atai, and Oller, Jr. 1974; Iino, Yabuta, and Thomas 2011; Kanzaki 2015; Leeming and Wong 2016). The other developed a measurement method of GEP based on dictation or read-aloud performances (Kotani and Yoshimi 2021a; Kotani and Yoshimi 2021b). Kotani and Yoshimi (2021a) and Kotani and Yoshimi (2021b) measured GEP by dictation performance and read-aloud performance, respectively. Therefore, it has not been examined to what extent the measurement performance can be improved by measuring GEP based both on dictation and read-aloud performances.

The goal of this study is to find an effective GEP-measurement method by comparing different patterns of sub-proficiencies. Hence, the research question is as follows.

- Which is the highest GEP-measurement performance among a dictation-based method, a read-aloud method, and a dictation and read-aloud-based method.

These three methods are compared not only with respect to the measurement accuracy, but also with respect to the easiness of measurement, that is, cost for the development of a method and the administration of the method.

The contribution of the present study includes (1) proposal of effective GEP-measurement method as a classroom-based assessment alternative to GEP tests, (2) empirical verification of GEP measurement methods, i.e., a dictation-based method, a read-aloud method, and a both dictation and read-aloud method, and (3) empirical verification of robustness of a GEP measurement method against English-language-class size of training data.

2 Related Research

2.1 GEP Measurement based on Dictation or Read-Aloud Performance

Kotani and Yoshimi (2021a) investigated the validity of dictation performance as an indicator of GEP by examining GEP measurement using non-linear regression analysis. The participants were 50 college students who dictated two texts that they were familiar with. The dependent variable was GEP in terms of TOEIC scores, and the independent variables included the dictation performance based on accuracy and the learners' subjective judgment, and linguistic features of the dictation texts. The measured GEP had a strong correlation with the observed GEP.

Kotani and Yoshimi (2021b) investigated the validity of read-aloud performance by examining GEP measurement using multiple linear regression analysis. The dependent variable was GEP in terms of TOEIC scores, and the independent variables included the read-aloud performance based on accuracy, speech rate, and the learners' subjective judgment, and linguistic features of the read-aloud texts. The measured GEP had a moderate correlation with the observed GEP.

2.2 Correlation Analysis between GEP and Dictation or Read-Aloud Performance

Previous research (Irvine, Atai, and Oller, Jr. 1974; Kazazoglu 2013; Wong and Leeming 2014; Leeming and Wong 2016; Yazdinejad and Zeraatpishe 2019) presumed that dictation performance is a good indicator of GEP, and compared learners' dictation performance scores with GEP scores such as the TOEIC and the TOEFL. The results showed a strong correlation between the dictation performance scores and GEP test scores. Hence, previous research has succeeded in demonstrating that dictation performance indicates test scores for GEP.

Coniam (1991) investigated the validity of speech rate in reading aloud as an indicator of GEP by examining the correlation between GEP and the speech rate. The participants were 83 secondary school students who read a short dialogue aloud (the length was uncertain). The read-aloud performance was evaluated on a seven-point Likert scale and included accuracy and fluency. GEP scores constituted reading-writing and speaking-listening scores. The speech rate was moderately correlated with read-aloud performance (r = 0.66), reading-writing performance (r = 0.55), listening-speaking performance (r = 0.60), and GEP (r = 0.59).

Iino, Yabuta and Thomas (2011) investigated the validity of read-aloud performance scores as an indicator of GEP by examining the correlation between GEP and read-aloud performance. In the read-aloud tests, 80 ESL learners read aloud four short texts. The results of the read-aloud tests were evaluated by three ESL teachers on a five-point Likert scale based on accuracy of pronunciation and accent, intelligibility of meaning units, and fluency of speech rate. The read-aloud scores had moderate correlation with test scores of GEP.

Lee (2014) investigated the validity of speech rate in reading aloud as an indicator of oral GEP by examining the correlation between oral GEP and speech rate. The participants were 46 college students who read aloud a short text. Oral GEP was evaluated by three ESL teachers and native English speakers. The oral GEP includes read-aloud performance and picture-cued storytelling. The speech rate had a moderate correlation with the read-aloud-based GEP and strong correlation with storytelling-based GEP.

3 Collection of Dictation and Read-Aloud Data

3.1 Participants

The participants of this study were 50 English learners. This number was determined to mimic a large English class that includes learners at different proficiency levels. This is because this study placed more emphasis on the practical application of model building than on the theoretical perspective. In addition, the participants were not randomly chosen. The use of class-size training data reveals the possibility of an ESL teacher to develop GEP measurement using training data compiled in the class.

Those who satisfied the following conditions participated in the experiment: their first language was Japanese; they were students of universities in the area where this study was carried out (28 men and 22 women; mean age, 20.8 years; standard deviation (*SD*), 1.3). The participants were paid a fee for participation.

3.2 Data Collection Procedures

Data instances to determine GEP comprised sentences transcribed by a learner, two types of dictation performance scores, speech sound pronounced by a learner, three types of read-aloud performance scores, five types of linguistic features extracted from reference sentences from a text material, and the learners' English test scores. The dictation and read-aloud data included 750 instances gathered from 50 learners' attempts to complete the text material consisting of 15 sentences.

The dictation task proceeded as follows: First, the 50 learners listened to sentences read aloud by a voice actor (woman, 35 years old) who was a native speaker of American English, and transcribed them sentence-by-sentence. Subsequently, the learners subjectively judged their ease of dictation (explained in Section 4.1).

The read-aloud task was performed as follows: First, the learners listened to a reference speech sound by the native speaker. Subsequently, they read a sentence aloud and subjectively judged the ease of reading aloud (explained in Section 4.2). Recording durations were collected to calculate speech rates.

Three instructions were given to the learners: 1) Each sentence could be listened or read twice if necessary; 2) Each task should be completed at a speed natural for the learner; and 3) It was forbidden to read fast or slowly, or to return and revise a sentence after moving on to the next sentence.

3.3 Text material

Two types of texts were selected from those distributed by the International Phonetic Association (1999) and Deterding (2006). As these texts include basic English sounds, an analysis of the learners' dictation and reading-aloud of these texts would reveal what types of English sounds influenced their listening and pronunciation.

These texts featured two of Aesop's Fables: The North Wind and the Sun (Text I) and The Boy Who Cried Wolf (Text II). Texts I and II contained five and ten sentences, respectively. It should be noted that Text I failed to encompass certain sounds, such as initial and medial /z/ and syllable initial /θ/. However, Text II included these missing sounds.

3.4 General English Proficiency

Learners' GEP was determined using their TOEIC Listening & Reading test scores, obtained in the current or previous year. The TOEIC Listening & Reading test was chosen, because the test scores had strongly correlated with GEP test results, that is, the Language Proficiency Interview developed at the Foreign Service Institute of U.S. Department of State (Educational Testing Service 1998), and this test has no dictation and read-aloud sections.

4 Features for Regression

This study measured GEP through regression based on dictation and/or read-aloud performance scores and the linguistic features of a sentence.

4.1 Dictation performance

The criteria for evaluating dictation performance comprised two indexes: learners' subjective judgment of their ease with dictation (EASE-D) and dictation accuracy (ACC-D).

EASE-D was scored using a five-point Likert scale for the learners' subjective judgment (1: easy; 2: somewhat easy; 3: average; 4: somewhat difficult; and 5: difficult). A lower EASE-D indicated that the learners judged the dictation to be easier.

ACC-D was calculated by dividing the Levenshtein edit distance between a given reference and a transcribed sentence with the number of characters in a longer sentence than the other. The Levenshtein edit distance reflects the differences between the two sentences due to the substitution, deletion, or insertion of characters. A lower ACC-D denoted that the learners completed the dictation more accurately.

4.2 Read-Aloud Performance

The criteria for evaluating read-aloud performance comprised three indices. They were learners' subjective judgment of the ease of reading aloud (EASE-R), read-aloud accuracy (ACC-R), and speech rate in words per minute (RATE-R).

EASE-R was determined by the learner's subjective judgment on a five-point Likert scale.

ACC-R was calculated by dividing the number of words correctly read aloud by the number of words in the corresponding sentence (0 indicated the absence of words correctly read aloud and 1 indicated that the learner read aloud all words correctly). Learners' reading aloud is evaluated word-by-word using a binary decision (correct or incorrect pronunciation) for constructing a naïve measurement method as a cost-effective method. The measurement accuracy might be improved if the reading aloud is evaluated phone-by-phone using the multiple decision for the appropriateness, but the development cost will increase. From the viewpoint of the cost-effectiveness, the reading aloud should be evaluated manually by an ESL teacher or automatically by a speech recognition tool, but we chose an English transcriber as an evaluator. A transcriber is supposed to provide strict evaluation for learners' reading aloud due to the unfamiliarity with learners' pronunciation. When this method is practically used in an ESL class, an ESL teacher should evaluate learners' reading aloud as strictly as possible to maintain the validity.

RATE-R was calculated by dividing the number of words by the duration of reading aloud.

4.3 Linguistic Features

In this study, linguistic features included sentence length, mean word length, number of multiple-syllable words, and word difficulty. These linguistic features were automatically derived from sentences in the text material.

Sentence length (Chall and Dial 1948) was defined as the number of words in a sentence.

The mean word length (Chall and Dial 1948) was derived by dividing the number of syllables by the number of words in the sentence. The number of syllables in a word (Stenton 2013) was counted using the following steps: count the vowels in the word, subtract any silent vowels, and subtract one vowel from every diphthong.

The number of multiple-syllable words in a sentence (Fang 1966) was derived using the formula $\sum_{i=1}^{N}(S_i - 1)$, where N denotes the number of words in the sentence and S_i denotes the number of syllables in the *i*-th word. This subtraction derivation ignores the single-syllable words.

Word difficulty (Kiyokawa 1990) was defined as the rate of words not listed in a basic vocabulary list in relation to the total number of words in the sentence.

The speech rate was defined as the number of words read aloud by the native speaker in one minute.

5 Measurement of GEP with Dictation and/or Read-Aloud Performances

Measurement methods were developed using support vector regression, considering GEP as the dependent variable. The independent variables were the dictation performance scores (EASE-D, ACC-D), the read-aloud performance scores (EASE-R, ACC-R, RATE-R), and the linguistic features.

GEP is measured by support vector regression. Support vector regression is less explainable but more accurate than multiple regression. Thus, an active feature of a learning model is not clearly described. Support vector regression was conducted using the function “svm()” defined in the “e1071” package of the software environment R (Meyer 2021). The radial basis function was set as a type of kernel function, and the other parameters of “svm()” were set as default.

The measurement methods were evaluated using a leave-one-out cross-validation test, considering one instance as test data and *n*-1 instances as training data. The training/test data included 750 instances.

Correlation analysis using *t*-test was performed between the measured and observed GEPs, where the significant threshold was set to 0.05. A statistically significant correlation was further examined using chi-square test for answering the research question, but a non-significant correlation was left out. The significance threshold was adjusted for multiple testing based on the false discovery rate (FDR) (Benjamini and Hochberg 1995).

To address the research question, three types of a measurement method were developed: one uses dictation performance scores, another read-aloud performance scores, and the other both dictation and read-aloud performance scores. In addition to each type of test scores, these methods use linguistic features of dictation/read-aloud materials. The research question was answered by testing the equality between the statistically significant correlation coefficients in the chi-square tests.

6 Experimental Results

6.1 Descriptive Statistics

Figure 1 shows the distribution of GEP. GEP followed a normal distribution according to the Kolmogorov-Smirnov test (K = 0.82, p = 0.25). The mean, minimum, and maximum GEP were 607.7, 295, and 900, respectively, and the *SD* was 184.45.

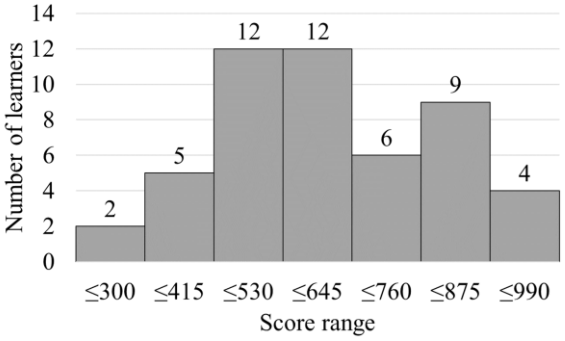


Figure 1: GEP distribution

Table 1 shows the means and SDs of the dictation and read-aloud performance scores, and Table 2 shows the means and SDs of the linguistic difficulty of sentences in the text material.

Performance score	*n*	*Mean*	*SD*
EASE-D	750	4.22	0.77
ACC-D	750	0.44	0.19
EASE-R	750	3.03	0.91
ACC-R	750	0.95	0.06
RATE-R	750	100.66	27.39

Table 1: Descriptive statistics of the dictation and read-aloud performances

Linguistic features	*n*	*Mean*	*SD*
Sentence length	15	21.93	7.57
Mean word length	15	1.26	0.11
Number of multiple-syllable words	15	5.93	2.84
Word difficulty	15	0.26	0.11
Speech rate	15	178.44	17.41

Table 2: Descriptive statistics of the linguistic difficulty of the sentences

6.2 Results and Discussion

Table 3 shows the correlation coefficients *r* between the measured and observed GEPs in the cross-validation tests. *Df* refers to the degree of freedom. D&R refers to a measurement method using dictation and read-aloud, D, a method using dictation, and R, the one using read-aloud. When the correlation coefficient was significantly different from zero, the coefficient was marked with an asterisk seen in all of the three types of measurement methods.

Measurement methods	*r*	*t*	*df*	*p*
D&R	0.80*	36.13	748	< 0.05
D	0.75*	31.17	748	< 0.05
R	0.59*	19.78	748	< 0.05

Table 3: Correlation coefficients of the three measurement methods

The scatterplots in Figure 2–4 show the correlations between the observed GEP and measured GEP (D&R, D, and R).

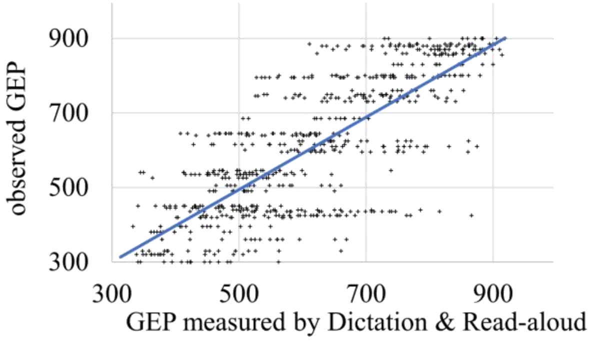


Figure 2: Scatter plot of GEP measured by D&R

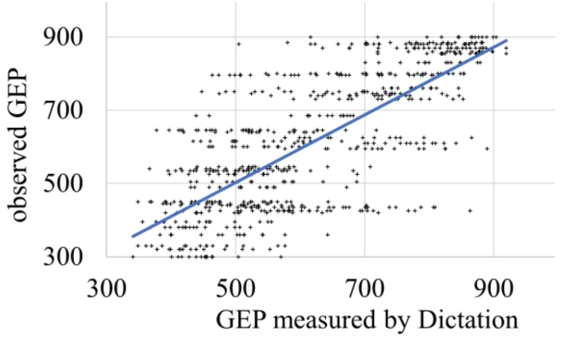


Figure 3: Scatter plot of GEP measured by D

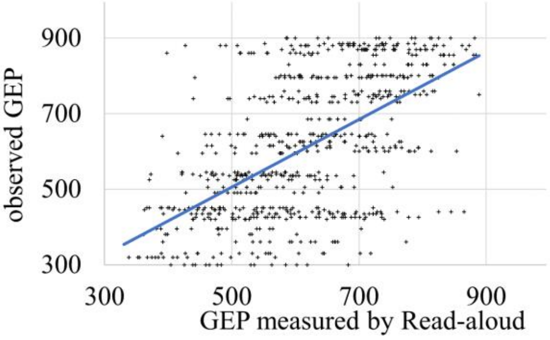


Figure 4: Scatter plot of GEP measured by R

Table 4 shows the results of the chi-square tests for equality of correlation among the three measurement methods. *FDR* refers to the significance threshold adjusted for multiple testing based on the false discovery rate. The chi-square value marked with an asterisk indicates significant differences between the correlation coefficients. The values of correlation coefficients are shown in a descending order: D&R > D > R. Table 4 indicated that statistically significance of pairs of correlation coefficients in the descending order: D&R > D, D&R > R, and D > R. The measurement method using D&R demonstrated the strongest correlation. That is, the result suggests that D and R are complementary to measure GEP.

Measurement methods	*chi-square*	*df*	*P*	*FDR*
D&R > D	4.89*	1	< 0.05	0.05
D&R > R	65.79*	1	< 0.02	0.02
D > R	34.78*	1	< 0.03	0.03

Table 4: Chi-square tests for equality among the three measurement methods

The significant difference in D > R can be explained in relation to association with TOEIC. Assuming that GEP can be measured with TOEIC scores, D > R indicated that TOEIC had stronger correlation with D than R. Both D and R share listening and reading comprehension skills, respectively. However, spelling in D and pronunciation in R are not examined in TOEIC. Hence, the correlation result, i.e., D > R, can be taken as a piece of evidence that spelling is more associated with TOEIC than pronunciation.

Therefore, the present study suggested that GEP should be measured with a method using D&R because of the strength of correlation, i.e., D&R > D > R. However, if an ESL teacher needs to decrease time for test administration and/or to reduce preparation tasks for test materials, a measurement can also be developed only with D instead of using D & R.

7 Conclusion

This study answered which of the GEP-measurement methods achieved the best performance among a dictation-based method, a read-aloud method, and a dictation and read-aloud-based method. Each method was developed by a non-linear regression analysis using dictation and/or read-aloud performance scores, and the linguistic features of the dictation/read-aloud materials. These methods were compared with respect to the measurement accuracy and the easiness of measurement.

The experimental result suggested that GEP should be measured with the dictation and read-aloud-based method, because the measured GEP had the strongest correlation with the observed GEP. However, if an ESL teacher needs to decrease testing time and/or preparation tasks for test materials, the diction-based method can also be utilized.

Future research should examine what combinations of dictation performances (EASE-D and ACC-D) and read-aloud performances (EASE-R, ACC-R, and RATE-R) can achieve the best measurement performance. It should also investigate how the measurement is dependent on learners' GEP.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers, 22300299, 15H02940, 17K18679, and 21K18373.

References

Yoav Benjamini and Yosef Hochberg.1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological), 57(1):289–300.

Jeanne S. Chall and Harold E. Dial. 1948. Predicting Listener Understanding and Interest in Newscasts. Educational Research Bulletin, 27(6):141–153+168.

David Coniam. 1991. Reading Aloud Speed as a Factor in Oral Fluency and General Language Proficiency? Hongkong Papers in Linguistics and Language Teaching, 14:47–69.

David Deterding. 2006. The North Wind versus a Wolf: Short Texts for the Description and Measurement of English Pronunciation. Journal of the International Phonetic Association, 36(2):187–196.

Educational Testing Service. 1998. TOEIC Technical Manuel. Educational Testing Service, Princeton: NJ.

Irving E. Fang. 1966. The Easy Listening Formula. Journal of Broadcasting, 11(1):63–68.

Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Automatic Assessment of English Proficiency for Japanese Learners without Reference Sentences based on Deep Neural Network Acoustic Models. Speech Communication, 116:86–97.

Atsushi Iino, Yukiko Yabuta, and Joel Thomas. 2011. Relationship between Criteria for Reading Aloud Evaluation and English Proficiency. Journal of the Chubu English Language Education, 40:159–166.

International Phonetic Association. 1999. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, Cambridge, UK.

Patricia Irvine, Parvin Atai, and John W. Oller, Jr. 1974. Cloze, Dictation, and the Test of English as a Foreign Language. Language Learning, 24(2):245–252.

Masaya Kanzaki. 2015. Minimal English Test: Item Analysis and Comparison with TOEIC Scores. Shiken: JALT Testing & Evaluation SIG newsletter, 19(2):12–23.

Semin Kazazoglu. 2013. Dictation as a Language Learning Tool. Procedia-Social and Behavioral Sciences, 70:1338–1346.

Hideo Kiyokawa. 1990. A Formula for Predicting Listenability: The Listenability of English Language Materials 2. Wayo Women's University Language and Literature, 24:57–74.

Katsunori Kotani and Takehiko Yoshimi. 2021a. Predicting English Proficiency with Read-Aloud Performance and Linguistic Difficulty of Sentences. Proceedings of International Technology, Education and Development Conference, pages 1180–1185.

Katsunori Kotani and Takehiko Yoshimi. 2021b. Prediction of General ESL Proficiency Considering Learners' Dictation Performance. The 3rd ETLTC International Conference on Information and Communications Technology, EDP Sciences, France.

Satsuki Kojima and Soichi Ota. 2012. Shadowing, Dictation and Reading Aloud: Which is Effective? Journal of The Japan Association of College English Teachers, 4:29–40.

Yongeun Lee. 2014. Quantifying English Fluency in Korean Speakers' Read-Aloud and Picture-Cued Storytelling Speech. Linguistic Research, 31(3):465–490.

Paul Leeming and Aeric Wong. 2016. Using Dictation to Measure Language Proficiency: A Rasch Analysis. Language Testing and Assessment, 5(2):1–25.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, Chih-Chen Lin. 2021. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Jose Noijons. 1994. Testing Computer Assisted Language Testing: Towards a Checklist for CALT. CALICO Journal, 12(1):37–58.

John W. Oller, Jr. 1983. Evidence for a General Language Proficiency Factor: An Expectancy Grammar. In John W. Oller, Jr. (Ed.), Issues in Language Testing Research, pages 3–10, Newbury House, Rowley: MA.

Henry L. Roediger III, Adam L. Putnam, and Megan A. Smith. 2011. Ten Benefits of Testing and their Applications to Educational Practice. In Jose P. Mestre and Brian H. Ross (Eds.), Psychology of Learning and Motivation: Cognition in Education, 55: 1–36, San Diego, CA: Elsevier Academic Press.

Anthony Stenton. 2013. The Role of the Syllable in Foreign Language Learning: Improving Oral Production through Dual-Coded, Sound-Synchronised, Typographic Annotations. Language Learning in Higher Education. Journal of the European Confederation of Language Centres in Higher Education, 2(1):145–161.

Ruslan Suvorov and Volker Hegelheimer. 2013. Computer-Assisted Language Testing. In Antony J, Kunnan (Ed.), The Companion to Language Assessment, John Wiley & Sons, Hoboken: NJ.

Elvis Wagner. 2020. Duolingo English Test, Revised Version July 2019, Language Assessment Quarterly, 17(3):300–315.

Aeric Wong and Paul Leeming. 2014. Using Dictation to Measure Language Proficiency. Language Education in Asia, 5(1):160–169.

Anoushe Yazdinejad and Mitra Zeraatpishe. 2019. Investigating the Validity of Partial Dictation as a Test of Overall Language Proficiency. International Journal of Language Testing, 9(2):44–55.

Curve-fitting of frequency distributions of dependency distances in a multi-lingual parallel corpus

Masanori Oya
School of Global Japanese Studies, Meiji University,
4-21-1, Nakano, Tokyo, Japan
masanori_oya2019@meiji.ac.jp

Abstract

This study discusses the curve-fitting results of the frequency distributions of dependency distances in sentences from a multi-lingual parallel corpus. It assumes that these distributions fit a mathematically well-defined distribution (the right truncated modified Zipf-Alekseev distribution) quite well, which indicates that these distributions of dependency distances reflect an aspect of the universal properties of natural language. However, the results of the curve-fitting of frequency distributions of the dependency distances of different dependency types do not demonstrate a suitable fit to the right truncated modified Zipf-Alekseev distribution, suggesting the necessity of further research.

1 Introduction

Dependency distance has been the center of focus of research on memory burden and syntactic complexity (Gibson, 1998, 2000; Gildea & Temperley, 2010; Grodner & Gibson, 2005; Li & Yan, 2021; Liu, 2007, 2008; Liu et al., 2017; Oya, 2013, 2021). The dependency distance between words in a dependency relation can be easily calculated; for example, in the sentence “Sarah has written an article in two months,” the noun *Sarah* depends on the auxiliary verb *has* as the subject, and the dependency distance between them is one. The noun *article* depends on *written* as its object, and the dependency distance between them is two.

Dependency distance has been argued as an aspect of the universal properties of natural languages; more specifically, shorter dependency distances are preferred to longer ones, possibly due to the upper bound of the short-term memory of humans (Gibson, 2000). It has been found out that the threshold of dependency distance is four across different natural languages (Liu, 2008; Oya, 2021). This means that the frequencies of dependency distances one, two, or three are much higher than the frequencies of dependency distances four or larger. When we plot the frequency distributions of dependency distances on an x-y plane with the x axis being the dependency distance and the y axis its frequency, then the graph has a long tail to the right.

In this context, frequency distributions of dependency distances have been discovered to fit the right truncated modified Zipf-Alekseev distribution (henceforth ZA distribution) quite well (Jiang & Liu, 2015; Liu, 2009; Ouyang & Jiang, 2017). The ZA distribution formula is illustrated below (Ouyang & Jiang, 2017; Popescu et al., 2014; Li and Yan, 2021):

$$y = cx^{a+b\ln x}$$

In the formula above, x is a dependency distance, y is its frequency, *c* is a constant, and the parameters *a* and *b* vary along with the fitness of the distributions of the frequencies. According to Li and Yan (2021), when the ZA distribution was fitted to the frequency distribution of dependency distances across essays written by Japanese EFL learners of different proficiencies, the results of good fitness were obtained.

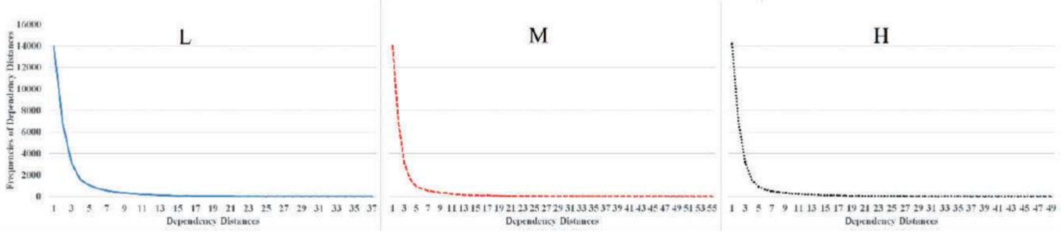


Figure 1. Frequency distributions of the dependency distances of Lower, Middle and High groups (Li & Yan, 2021, p.180)

Li and Yan (2021) also found out that learner proficiencies are reflected in the different parameters of the ZA distribution. That is, parameter *a* increases, and parameter *b* decreases, from the lower (L in the table below), the intermediate (M), and to the higher proficiency learner groups (H).

Group	a	b	n	α	X^2	$P(X^2)$	DF	C	R^2
L	0.66	0.4634	45	0.4646	366.69	0	40	0.0112	0.9969
M	0.6982	0.4056	85	0.4605	611.47	0	60	0.0201	0.9944
H	0.7206	0.3898	63	0.4608	805	0	55	0.0261	0.9929

Table 1. Fitting ZA distribution to the frequency distributions of dependency distances of different groups (Li & Yan, 2021, p.181)

2. Background of this study

This study follows Li and Yan’s (2021) research, attempts to fit the ZA distribution to the frequency distribution of dependency distances in sentences taken from a multi-lingual parallel corpus, and hopes to find aspects of properties that are universal across different languages. More specifically, we focus on the similarities/differences of parameters *a* and *b* across different languages in a multi-lingual parallel corpus, with the assumption that such comparisons can indicate the similarities/differences of these languages mathematically, with reference to the ZA distribution. If the parameters *a* and *b* of one language are found to be close to those of another language, it will indicate their similarity as far as their fitness to the ZA distribution is concerned. If, on the other hand, these parameters are widely different across the two languages, it will be indicative of their difference in terms of the ZA distribution.

Along with this distribution-fitting of individual languages, we also focus on the distributions of dependency distances of different dependency types, e.g., the distribution of dependency distances between a verb and its subject, and its attempt to fit to the ZA distribution. Similar to the curve-fitting of the distributions of dependency distances across all dependency types, we focus on the similarities/differences of parameters *a* and *b* of different languages in the same multi-lingual parallel corpus. This fitting is expected to reveal the behaviors of the dependency distances of different types, and to open up the possibility of the investigation of a certain aspect of the universal properties of natural languages from a more fine-grained perspective.

This study poses the following research questions:

1. Will the frequency distribution of the dependency distances in a multi-lingual parallel corpus fit suitably to the ZA distribution?
2. Will similarities/differences of languages be reflected on the parameters of the ZA distribution?
3. Will the frequency distribution of the dependency distances of different dependency types be reflected in the parameters of ZA distribution?

2.1 Data

This study uses *Parallel Universal Dependencies Treebanks 2.7* (henceforth PUD) to answer the abovementioned research questions. These treebanks were created for the purpose of the shared task on Multilingual Parsing from Raw Text to Universal Dependencies at CoNLL 2017 (http://universaldependencies.org/conll17/). PUD is a parallel corpus consisting of 20,000 sentences with aligned translation pairs across 20 languages. (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Icelandic, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish). Of the 1,000 sentences of each language, 750 were those which had been translated directly from English texts, and the remaining 250 were translated from German, French, Italian, or Spanish; these sentences had been first translated into English, and then further translated into each of these languages. The translation of these sentences was conducted manually by professional translators, and these sentences were further annotated with morphological and syntactic tags by Google. These annotations are converted into UD (Universal Dependencies) by UD community members, according to UD Ver. 2 guidelines. The UD website elucidates further details (https://universaldependencies.org/).

2.2 Procedure

We can calculate the dependency distance between two words in dependency relationship. We can obtain the dependency distance of each dependency relation in PUD through the dependency tag on every word in a sentence. It is conducted by an original Ruby script which was the same used in Oya (2021). On the basis of the output of the script, we obtained the frequencies of the dependency distances of all the dependencies in the sentences in each of the languages in PUD, as well as the frequencies of dependency distances of different dependency types. This study focuses on the frequencies of the dependency distances of the dependency types *nsubj* (dependency between a verb and its nominal subject), *obj* (dependency between a transitive verb and its direct object), *amod* (dependency between a noun and an adjective modifying the noun), and *advmod* (dependency between an adverb and a noun modified by the adverb). Then, the frequency distributions of dependency distances are calculated, and fitted to the ZA distribution by means of the Altmann-fitter v.3.1.0 (http://www.ram-verlag.biz/altmann-fitter/), which is the same application that Lin and Yan (2021) used to curve-fit their data.

2.3 Results

The table below illustrates the results of the curve-fitting of the dependency distances' frequency distributions in the 20 PUD languages to the ZA distribution. The R^2 values of all these languages (except for French) are above 0.98, which indicates that they all fit the ZA distribution quite well.

	a	*b*	n	α	X^2	$P(X^2)$	C	DF	R^2
Arabic	0.9274	0.2634	51	0.5214	244.27	0	0.0118	46	0.9985
Czech	0.3531	0.4618	45	0.4061	211.4	0	0.0114	40	0.9971
German	0.2154	0.4234	51	0.3583	411.76	0	0.0193	46	0.9925
English	0.206	0.5165	56	0.3638	490.23	0	0.0232	48	0.9924
Spanish	0.4536	0.4437	59	0.407	995.06	0	0.0427	52	0.9853
Finnish	0.3764	0.4738	44	0.4245	96.4	0	0.0061	39	0.9998
French	0.1188	0.5258	52	0.4063	1491.42	0	0.0603	47	0.9787
Hindi	0.5507	0.2748	53	0.4605	1050.47	0	0.0441	48	0.9915
Indonesian	0.6206	0.3706	46	0.4825	224.7	0	0.0116	41	0.9976
Icelandic	0.3406	0.4987	49	0.4317	290.77	0	0.0154	42	0.9957
Italian	0.2291	0.5015	63	0.4047	1263.5	0	0.0532	52	0.9816
Japanese	1.4917	0.1277	72	0.4327	1486.15	0	0.0516	67	0.9894
Korean	0.8252	0.2137	46	0.5528	527.16	0	0.0318	41	0.9967
Polish	0.5594	0.3939	42	0.4628	194.13	0	0.0106	37	0.9976
Portuguese	0.2558	0.4947	61	0.4058	1097.69	0	0.0469	51	0.9836
Russian	0.5218	0.4262	46	0.4259	384.96	0	0.0199	41	0.9936
Swedish	0.1053	0.5602	50	0.3888	537.5	0	0.0282	42	0.9903
Thai	0.5959	0.4327	42	0.5327	101.95	0	0.0046	37	0.9994
Turkish	0.7389	0.2438	40	0.5198	552.58	0	0.0327	35	0.9954
Chinese	0.7587	0.2628	47	0.4085	181.6	0	0.0085	42	0.9995

Table 2. The results of the curve-fitting of the frequency distributions of the dependency distances in the 20 languages in the PUD, to the ZA distribution.

The figure below illustrates that, when these languages are plotted according to their parameters *a* and *b*, they demonstrate a linear distribution; $R^2 = .82$, $p < .01$. In this linear distribution, we can notice that languages of the Indo-European family seem to form one cluster in which the parameter *a* is less than 0.6 and the parameter *b* is within the range of 0.4 and 0.6 (except for Hindi), while non-Indo-European languages form another cluster in which the parameter *a* lies within the range of 0.6 and 1 and the parameter *b* is less than 0.4 (except for Finnish and Thai). Japanese stands out from these two clusters.

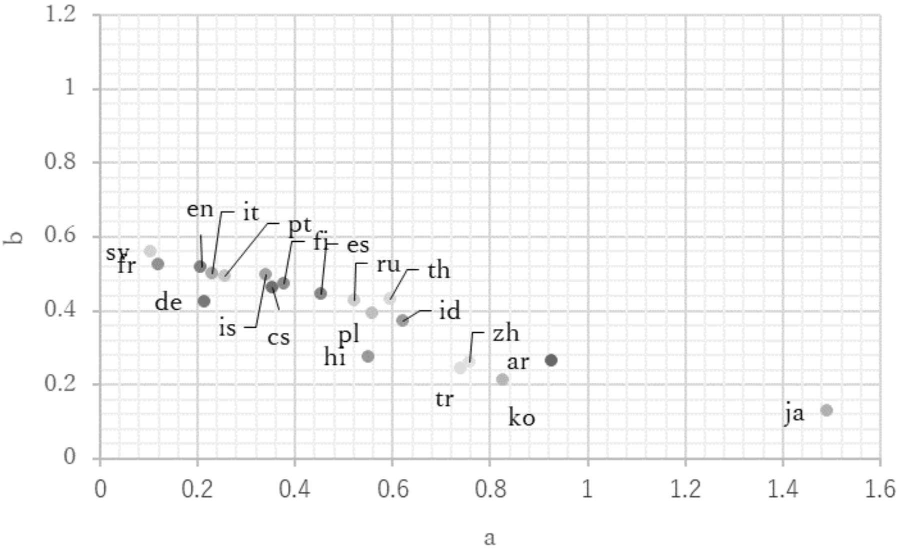


Figure 3. The plots of parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of all dependency types in the PUD are fitted to the ZA distribution

Unlike the case of the plot above which takes all the dependency types into consideration, when the dependency distances of *nsubj* in the PUD are fitted to the ZA distribution, there is a weak linear distribution of the languages in the PUD in terms of their *a* and *b* parameters of the ZA distribution; $R^2 = .27$, $p < .02$.

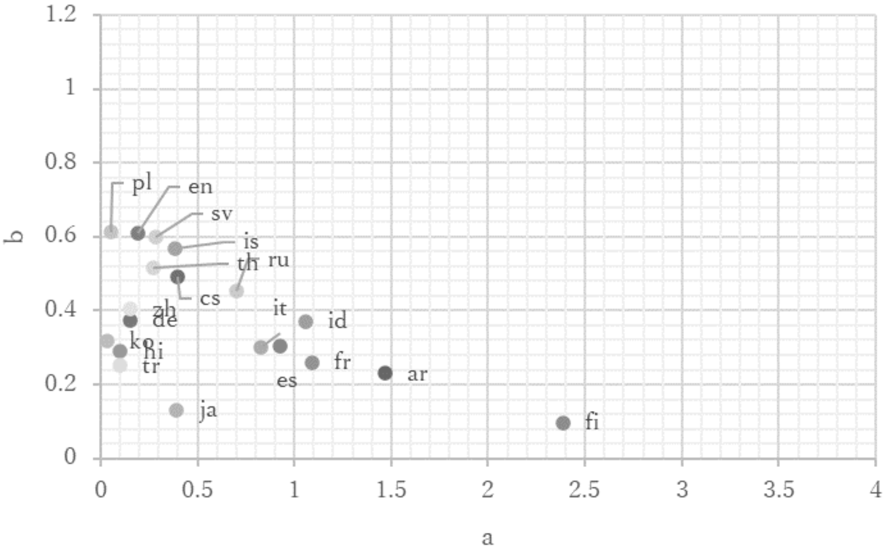


Figure 4. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *nsubj* in the PUD are fitted to the ZA distribution

When the dependency distances of *obj* in the PUD are fitted to the ZA distribution, there is no linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; $R^2 = .13$, $p < .12$.

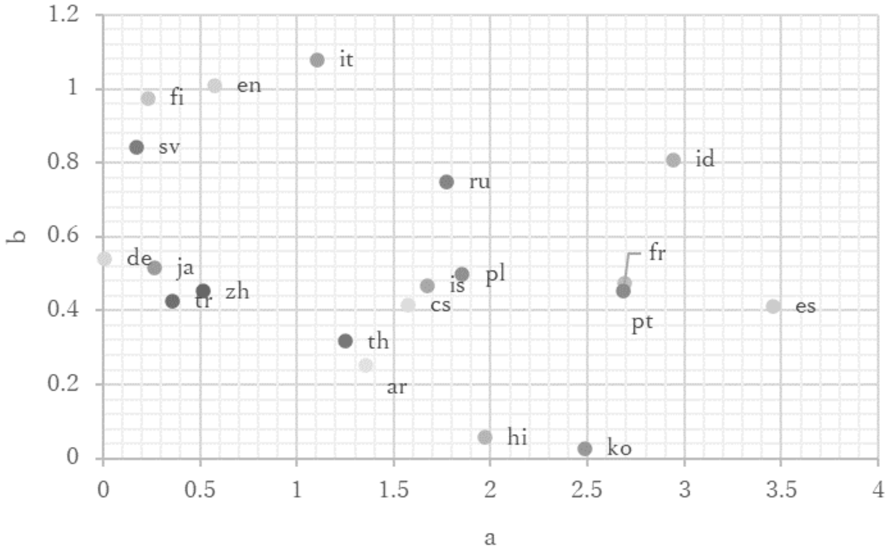


Figure 5. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *obj* in the PUD are fitted to the ZA distribution

When the dependency distances of *amod* in the PUD are fitted to the ZA distribution, there is no linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; $R^2 = .08$, $p < .21$. This non-linear distribution is partly due to the fact that the frequencies of *amod* in Arabic, Indonesian and Japanese do not fit the ZA distribution. When these frequencies are excluded, the distribution turns out to be linear; $R^2 = .67$, $p < .01$.

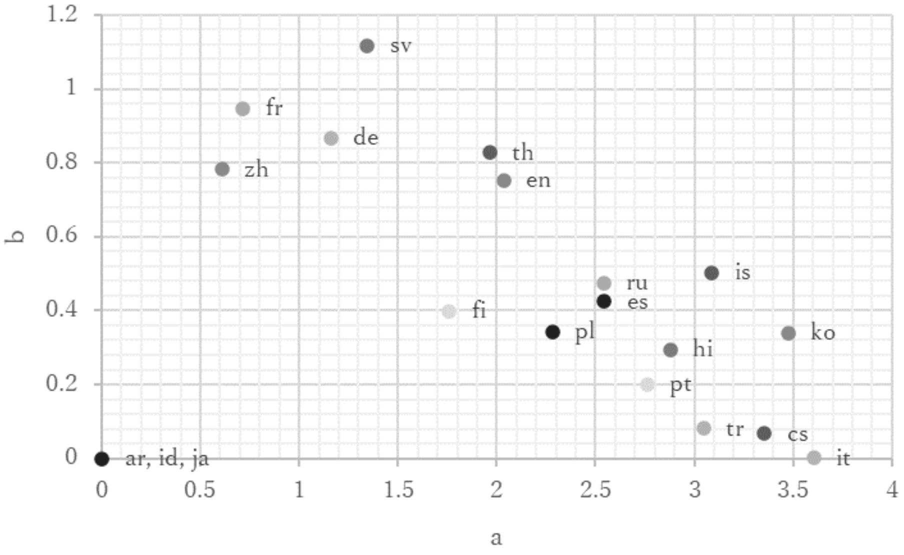


Figure 6. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *amod* in the PUD are fitted to the ZA distribution

When the dependency distances of *advmod* in the PUD are fitted to the ZA distribution, there is a weak linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; $R^2 = .46$, $p < .01$.

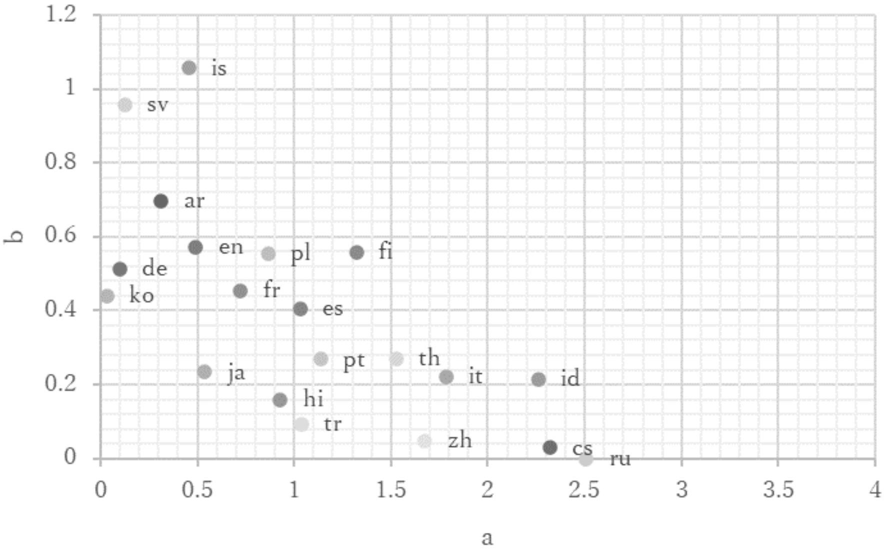


Figure 7. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *advmod* in the PUD are fitted to the ZA distribution

3. Discussion

The results answer R.Q.1 and R.Q.2, and indicate that the frequency distribution of dependency distances in a multi-lingual parallel corpus fit well to the ZA distribution (R.Q.1), and that differences between languages are reflected in the different parameters of the ZA distribution (R.Q. 2). The clusters of parameters seem to reflect the differences of language families they belong to; however, further investigations that include more languages which have not been included in PUD are necessary.

The aforementioned results do not seem to answer R.Q.3 positively. The frequency distributions of the dependency distances of different dependency types in the multi-lingual parallel corpus do not necessarily fit well with the ZA distribution. These results may be due to the fact that the same dependency type (e.g., *obj*) can be used differently in different languages, even in a parallel corpus in which sentences of different languages are aligned according to their semantic parallelism; for example, the direct object of a verb in one sentence of a language can be expressed not as the direct object of the sentence's translation in another language, and vice versa. The same is true for other dependency types. This may result in more random distribution of dependency distances, and less linear distribution of the parameters *a* and *b*.

We can notice different degrees of linear distributions of the parameters *a* and *b* across different dependency types. For example, the parameters *a* and *b* of the dependency type *nsubj* seems to have relatively higher degree of linear distribution than other dependency types. This may reflect the fact that what is expressed as the subject of a verb in one language is often expressed as the subject of a verb in other languages, and also their dependency distances (and their frequencies) are similar with each other. As such, we still need further investigation into different distributions of dependency distances of different dependency types across different languages, and finding language-(in)dependent patterns across these differences remains to be one of the goals in future research.

4. Conclusion

This study reported the results of the curve-fitting of frequency distributions of dependency distances in sentences within a multi-lingual parallel corpus. It was found that these distributions fit the ZA distribution fairly well. This indicates that these distributions of dependency distances can reflect a certain aspect of the universal properties of natural language. However, the results of the curve-fitting of frequency distributions of the dependency distances of different dependency types do not demonstrate a suitable fit to the ZA distribution. Thus, these results require us to further investigate the behaviors of different types of dependencies in terms of their distances and their frequencies, both within the same parallel corpus and other types of corpus data.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20K00583.

References

Marie-Catherine de Marneffe, Bill MacCartney and Christopher Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. LREC *2006*.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre and Dan Zeman. 2021. Universal Dependencies. Computational Linguistics, 47 (2): 255-308. https://doi.org/10.1162/COLI_a_00402.

Richard Futrell, Kyle Mahowald and Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. Proceedings of Natural Academy of Science, 112(33):10336-10341. https://doi.org/10.1073/pnas.1502134112

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95-126). MIT Press, Massachusetts, US.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? Cognitive Science, 34(2):286-310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science, 29(2):261-290. https://doi.org/10.1207/s15516709cog0000_7

Shin-Ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Learner corpus studies in Asia and the world, 1: 91-118.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. Language Sciences, 50: 93-104. https://doi.org/10.1016/j.langsci.2015.04.002Jiang, J. & H.

Jingyang Jiang and Haitao Liu (eds). 2018. Quantitative Analysis of Dependency Structures. Berlin: De Gruyter Mouton.

Jingyang Jiang and Jinghui Ouyang. 2017. Dependency distance: A new perspective on the syntactic development in second language acquisition. Physics of Life Review, 21: 209-210.

Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. Journal of Quantitative Linguistics 27(1): 62-79.

Wenping Li and Jianwei Yan. 2021. Probability distribution of dependency distance based on a Treebank of Japanese EFL learners’ Interlanguage. Quantitative Linguistics, 28(2): 172-186, DOI: 10.1080/09296174.2020.1754611

Haitao Liu. 2007. Probability distribution of dependency distance. Glottometrics, 15:1-12.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science, 9(2):159-191.

Haitao Liu. 2009. Probability distribution of dependencies based on Chinese dependency treebank.

Journal of Quantitative Linguistics, 16(3): 256-273. https://doi.org/10.1080/09296170902975742

Haitao Liu, Chunshan Xu and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 21:171-193. https://doi.org/10.1016/j.plrev.2017.03.002

Jinghui Ouyang and Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? Journal of Quantitative Linguistics, 25(4): 295-313. https://doi.org/10.1080/09296174.2017.1373991

Masanori Oya. 2013. Degree centralities, closeness centralities, and dependency distances of different genres of texts. Selected papers from the 17th Conference of Pan-pacific Association of Applied Linguistics, 42-53.

Masanori Oya. 2021. Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus. Pacific Asia Conference on Language, Information and Computation (PACLIC) 35, Nov 7, 2021

Ioan-Iovitz Popescu, Karl-Heinz Best and Gabriel Altmann. 2014. Unified modeling of length in language. Studies in Quantitative Linguistics 16. RAM-Verlag.

Yalan Wang. 2020. Quantitative Analysis of Dependency Structures. Journal of Quantitative Linguistics 27(1): 83-91.

Dan Zeman. 2015. Slavic Languages in Universal Dependencies. Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning. Slovakia.

Dan Zeman and Joakim Nivre, et al. 2020. Universal Dependencies 2.7, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3424.

Measuring public opinion on the import of US pork in Taiwan

Yu-Lun Huang
National Chengchi University, Taiwan
110753134@nccu.edu.tw

Yu-Yun Chang
National Chengchi University, Taiwan
yuyun@nccu.edu.tw

Jyi-Shane Liu
National Chengchi University, Taiwan
liujs@nccu.edu.tw

Chang-Yi Lin
National Taiwan Normal University, Taiwan
61015003E@gapps.ntnu.edu.tw

Abstract

As the import of pork with ractopamine from the United States (US) has stimulated a heated discussion on social media, the purpose of this exploratory study is to assess public opinion on the import of US pork by examining the interplay between the two sub-issues, namely food security (FS) and political economy (PE), at the word, sentence, and time series levels. First, we collect comments related to the US pork issue on PTT, the largest online discussion forum in Taiwan, throughout the year 2021. Second, at the word level, we apply Latent Dirichlet Allocation (LDA) to extract the main topics. Third, at the sentence level, we sample a subset of data for manual annotation. Taking the annotated subset as the gold standard, we use the pretrained BERT-base-Chinese model for the downstream topic classification task to predict the whole dataset. Fourth, at the time series level, we employ Detrended Fluctuation Analysis (DFA) to evaluate the predicted number of comments for FS and PE sub-issues. Two main results are obtained: (1) the theme of political economy interweaves with the concept of food safety at both word and sentence levels; (2) DFA shows the absence of a more salient perspective between the FS and PE sub-issues at the time series level.

1 Introduction

In Taiwan, social issues are typically divided into different discursive dimensions according to political ideology. Take the US pork issue as an example, the Democratic Progressive Party's (DPP) decision to lift the restriction on the import of pork with ractopamine was made to strengthen Taiwan's economic ties with the US, while the Kuomintang Party's (KMT) opposition to this policy heightens the potential risk to food safety and facilitates the holding of a national referendum banning the US pork importation. As the referendum is intended to represent the views of citizens for involvement in the policy-making process, the question then arises as to which perspective the political parties attempt to shape public opinion from. As social media mining has become a powerful tool to investigate the formation and change of public opinion (Anstead & O'Loughlin, 2015), this study evaluates the evolution of public views via the Professional Technology Temple (PTT), Taiwan's largest online bulletin board where the issue of US pork has been extensively discussed.

In their study of stance detection of comments posted on PTT, Chuang and Hsieh (2015) examine the cross-fertilization of ideas between netizens participating in online discussions and audiences reading the comments. Indirect participation in the conversations notwithstanding, the audiences still receive the messages conveyed by the interlocutors. Hence, discussion on PTT represents a dominant force shaping public attitudes toward a particular issue. However, studies to date on stance detection have frequently drawn on social media data at a document-level where each text was written by a single author (e.g., Faulkner, 2014; Walker et al., 2012). In fact, short comments posted below these articles are opinion-rich resources which influence the attitudes of netizens. This paper engages with the tendency of public opinion reflected in short comments provided by multiple authors.

Tu et al. (2021) analyze the issue of US pork from a media and communication perspective, exploring the behavioral patterns of netizens, such as posting time and posting frequency. However, they (2021) neglect the linguistic aspects in media discourse and mention without further elaboration that the US pork issue is recognized as both a food safety concern and a politico-economic issue. Building on Tu et al.'s (2021) study and taking it as a point of departure, the current research aims to examine public attitudes toward the US pork issue from a linguistic standpoint and measure the extent to which the two sub-issues, i.e., food security (FS) and political economy (PE), intersect and interact with each other through a quantitative lens.

In this study, topic modeling is conducted to retrieve linguistic information about the US pork issue at three different levels: word, sentence, and time series. First, at the word level, we adopt Latent Dirichlet Allocation (LDA), a topic model which has been extensively used to build clusters of documents, to extract the main topics. While LDA provides a useful lens for topic modeling, the topics extracted by LDA only model correlations among words without explicitly representing correlations among the topics (Li & McCallum, 2006). Therefore, at the sentence level, a semi-supervised transfer learning method based on the pretrained BERT model is applied to characterize the interaction between the FS and PE sub-issues by incorporating contextual information. Finally, at the time series level, we take the time axis into account by using Detrended Fluctuation Analysis (DFA) to evaluate the evolution of public views. The proposed method seeks to contribute to research on social media mining by assessing the interplay between sub-issues of a specific social issue at the word, sentence, and time series levels.

The remainder of this paper is organized as follows. It first reviews existing literature on topic modeling and time series data analysis (section 2). Then it presents data collection, the annotation scheme, and the use of LDA, the pretrained BERT model, and DFA (section 3). After that, it not only illustrates the word clouds generated by LDA and presents the predicted results obtained by the BERT model but also reveals the dynamics of public opinion on the US pork importation and demonstrates the DFA results (section 4). This paper concludes with a brief discussion of the significance of our results and implications for understanding the chronological development of public concern over a particular social issue (section 6).

2 Related Work

Prior studies pertaining to topic classification have typically utilized machine learning algorithms. One line of research has focused on unsupervised learning techniques in which words are clustered into a set of topics, such as LDA (e.g., Hong & Davison, 2010) and Latent Semantic Analysis (e.g., Valdez et al., 2018). Another line of research has concentrated on supervised learning algorithms, such as Naive Bayes, Support Vector Machine, Logistic Regression, and Decision Tree (e.g., Lee et al., 2011; Wang & Manning, 2012). In addition to traditional machine learning models, deep learning models, such as LSTM (Hochreiter & Schmidhuber, 1997) and BERT (Devlin et al., 2018), have received considerable attention in the field of natural language processing in recent years. However, due to high computation cost for training large models with numerous parameters, a transfer learning method using pre-trained models is widely employed to perform text classification, such as topic detection (e.g., Houlsby et al., 2019; Yin et al., 2019). In the current study, we present a complementary approach using both unsupervised and semi-supervised transfer learning for the investigation of public opinion on the import of US pork.

2.1 Latent Dirichlet Allocation

One promising approach to unsupervised topic modeling is the implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a valuable tool for automated text summarization and visualization. On the basis of the Bag-of-Words paradigm, LDA is a generative probabilistic model that describes each document as a distribution of topics and each topic as a distribution of words. To generate a document, LDA first samples a per-document multinomial distribution over topics from a Dirichlet distribution. Then it repeatedly samples words in each document from the corresponding multinomial distribution; the procedure is repeated for all the words in each document.

2.2 Topic classification using BERT

As Devlin et al. (2018) elaborate, during the pre-training stage, the BERT-base-Chinese model is trained on two tasks, masked language modelling and next sentence prediction, to learn from the surrounding contextual information in the Chinese Wikipedia corpus. In the fine-tuning stage, the ⟨CLS⟩ token at the beginning of each sequence, whose embeddings are used as the representation of each input sequence, is fine-tuned for task-specific objectives. In this study, we annotate a subset of data and employ a transfer learning paradigm based on the pretrained BERT-base-Chinese model to solve the downstream topic classification task. After BERT makes predictions on the whole dataset, we obtain the predicted number of comments throughout the year 2021.

2.3 Detrended Fluctuation Analysis

Since the concept of timeline is integrated into the trend of public opinion, we apply the DFA to the time series analysis of the predicted number of comments in order to evaluate the changes in topics over time. More specifically, we quantify the scaling properties of time series movements in public opinion. The DFA is a useful technique for the detection of long-range correlations embedded in non-stationary time series (Peng et al., 1995; Kantelhardt, 2011). It is composed of the following steps: First, generate an integrated time series of N data points $\{x_1, x_2, \ldots, x_N\}$:

$$X(i) = \sum_{k=1}^{i} x_k - \langle x \rangle$$

where $\langle x \rangle$ is the mean of the above series.

Second, divide the integrated time series into $N_s = int(N/s)$ non-overlapping windows of equal length, each consisting of s number of samples.

Third, calculate the least square fit of preferred order to the data points in each window. This represents the local trend p_n. Next, subtract the local trend from the corresponding window to calculate the detrended series for windows of size s.

$$X_s(i) = X(i) - p_n(i)$$

where $p_n(i)$ is the fitting polynomial in window n, for each window $n = 1, 2, \ldots, \mathrm{N_s}$.

Fourth, determine the variance of the detrended series in each window n by averaging over all data points:

$$F_s^2(n) = \langle X_s^2(i) \rangle = \frac{1}{s} \sum_{i=1}^{s} X_s^2\, [(n-1)s + i]$$

then, average over all segments and take the square root to compute the fluctuation function $F(s)$:

$$F(s) = \left[\frac{1}{N_s} \sum_{n=1}^{N_s} F_s^2\,(n) \right]^{1/2}$$

$F(s)$ increases with the growing value of s. If the data are long-range power-law correlated, then $F(s)$ will show in the form of a power-law:

$$F(s) \propto s^{\propto}$$

where $\propto$ is the scaling exponent of the time series. The case $0.5 < \propto < 1$ for all scales s confirms the presence of long-range correlations in a time series (Kantelhardt, 2011). The greater the value $\propto$ is, the higher the correlations in the signal are.

To the best of our knowledge, the literature on time-series measurement of public opinion is rather sparse, particularly Chinese textual data, in that previous relevant studies are predominantly from English-speaking countries (e.g., Brulle et al., 2012; Stimson, 1999). This study is expected to pave the way for a novel data-driven analytical framework for applying machine learning and time-series analysis to social media content.

3 Methodology

The methodology is made up of four phases. First, we collect comments related to the US pork controversy posted on PTT. Second, at the word level, we perform topic modeling using LDA and text visualization using word clouds. Third, at the sentence level, we annotate the dataset and then present a transfer learning method using the pretrained BERT model for the downstream topic classification task. Fourth, at the time series level, we take the results predicted by the BERT model and conduct the DFA to assess the public perceptions of the US pork importation.

3.1 Data collection

We collect a total of 74,040 comments posted in the year 2021 regarding the US pork importation from the Gossiping Board[1] and HatePolitics Board[2] on PTT according to Tu et al.'s (2021) finding that articles relevant to the import of US pork are constantly posted on these two boards to stimulate discussion. Each comment is required to satisfy two criteria in order to be included. First, the data are collected between January 1 and December 31, 2021. Moreover, to conduct time-series analysis of the evolution of public views, each comment is matched with its corresponding date. Second, on the basis of Tu et al.'s (2021) study, each comment must contain at least one of the six keywords listed below: 美豬 (US pork), 美國豬肉 (pork from the US), 萊豬 (pork containing ractopamine), 萊克多巴胺 (ractopamine), 萊劑 (ractopamine), and 瘦肉精 (leanness-enhancing feed additive). This data extraction method confirms all selected comments are relevant to the US pork issue.

Subsequent to data preprocessing, a proportion of the whole dataset is randomly sampled to construct a subset of 15,190 comments for manual annotation. Table 1 demonstrates the size of the corpus used in this research.

	Subset	All
# Comments	15,190	74,040
# Tokens	137,007	662,487
# Word types	14,806	47,666

Table 1. Corpus information.

3.2 Procedure

Firstly, at the word level, we set the number of clusters to be 2 based on a preliminary analysis of the top 200 keywords extracted by the TF-IDF method. Despite the problem of selecting the optimal number of topics, from the examination of the FS and PE sub-issues, we seek to investigate whether the topic keywords extracted by LDA are clustered together according to the relation between a food security standpoint and a politico-economic point of view.

Then, we exploit LDA to cluster comments on PTT into two topics and obtain 300 extracted keywords to represent each topic after removing stop words and expressions referring to the import of US pork into Taiwan (e.g., 美豬 US pork, 美國豬肉 pork from the US, 萊克多巴胺 ractopamine, 萊豬 pork with ractopamine, 台灣 Taiwan, and 進口 import). Through an analysis of the extracted topic keywords, we note that the FS and PE sub-issues are difficult to separate because the keywords representing the two sub-issues intersect with one another; thus, during manual labeling, we divide the comments into four categories by including two more categories (Both and Other).

Secondly, at the sentence level, we manually annotate the subset by classifying each comment into four categories: FS, PE, Both, and Other. To reduce the burden of manual annotation, we present a semi-automatic corpus-based approach inspired by Liu (2012), in which TF-IDF, syntactic rules, and co-occurrence patterns are applied to produce the keyword lists. Nonetheless, due to the inevitability of errors caused by semi-automatic methods, a subsequent manual editing process is implemented by two well-trained annotators with linguistic background to ensure annotation quality.

Next, we adopt a transfer learning method in which the annotated subset is taken as the input to fine-tune the pretrained BERT-base-Chinese model (with 12-layer, 768-hidden, 12-heads, and 110M parameters) for the downstream topic classification task. During this phase, we divide the annotated subset into 80% training set and 20% test set. Then, our BERT model is trained for 20 epochs with a batch size of 64 and a learning rate of 5e-5. After the fine-tuning stage, our BERT model makes predictions on the whole dataset. Then, we count the predicted results to estimate the frequency distribution of the number of comments classified into the four categories per day in the year 2021, plotting the frequency distribution to show the evolution of public attitudes toward the US pork importation.

Finally, at the time series level, we apply DFA to the predicted frequency distribution so as to confirm the existence of long-range correlations in time series data.

[1] https://www.ptt.cc/bbs/Gossiping/index.html

[2] https://www.ptt.cc/bbs/HatePolitics/index.html

3.3 Annotation scheme

The annotation guidelines, listed in Table 2, are formulated to ensure accuracy and consistency of manual annotation. Chief among the annotation scheme is the target of the opinion as a mechanism to determine whether a comment is categorized as Food Safety or Political Economy; a comment is classified as FS if the target of the opinion is either the US pork or ractopamine, whereas a comment is classified as PE if the target of the opinion is the issue of US pork. However, there are a sizeable proportion of vague comments posted on PTT which express ambiguous meanings (Chuang & Hsieh, 2015) and are thus grouped into the Other category, including (1) vague expressions, such as the phrase 塔綠班 (tǎlyùban), a near homophonic pun on 塔利班 (Taliban); (2) truncated comments resulting from the word count limit of 27 Chinese characters; (3) vague messages which can only be interpreted as meaningful when processed together with the contextual information given by previous comments.

Category	Definition and example
Food Safety	The target of the opinion is either the US pork or ractopamine
	(e.g.) 有安全的豬肉為什麼要吃萊豬 ‘Why bother eating pork with ractopamine at all when we have safe pork to eat?’
Political Economy	The target of the opinion is the US pork issue
	(e.g.) 萊豬本來就是外交和國際政治的議題啊 ‘Pork with ractopamine has been a diplomatic and international political issue.’
Both	The comment satisfies the criteria of both Food Safety and Political Economy
	(e.g.) 進萊豬卻又不准賣場標示萊豬，有事嗎 ‘(The government allowed) the import of pork with ractopamine but banned the market from labeling it, is there a problem?’
Other	The topic of the comment is (1) related neither to Food Safety nor to Political Economy or (2) vague and ambiguous
	(e.g.) 把萊豬加個 i 就變潮啦 i 萊豬 ‘If we add an “i” before pork with ractopamine, it will become popular! iPork with ractopamine!’

Table 2. Annotation guidelines.

Besides the vagueness of language categorized into Other, we discuss the challenges annotators experience during the coding process. First, annotators have difficulties when deciding whether a comment that contains the word 吃 (eat) can be directly labeled as FS. In example (1), the phrase 吃虧 (being shortchanged) includes the word 吃 (eat), a keyword in the FS list. Considering this phrase along with 貿易組織 (trade organization), our semi-automated approach assign this comment with the Both pre-defined label; however, this comment does not convey the meaning of eating US pork, so the annotators made a correction by changing the label to PE. As for example (2), the phrase 進入校園 (introducing into the campus) implies that the author of the comment regards the US pork issue as a food security concern, because the author questions the safety of US pork with ractopamine. In example (3), “eating the US pork containing ractopamine” is interpreted as FS, whereas “breathing polluted air” is labeled as an issue of PE because of the author’s ironic tone against the government’s failure to address the problem of air pollution. Therefore, considering the whole sentence, example (3) is annotated as Both. Finally, example (4) is originally classified as FS through our semi-automated method due to the presence of the verb 吃 (eat), yet the agent of the verb is Taiwanese pigs rather than Taiwanese people. Since the food safety issue focuses on whether the importation of US pork has a harmful effect on the health of Taiwanese people rather than Taiwanese pigs, example (4) is neither related to the politico-economic dimension issue nor to the food security crisis and is thus categorized as Other.

(1) 欸進萊豬是我們吃虧 怎不先讓我們進貿易組織
We are shortchanged on the import of US pork containing ractopamine, so why not first allow us to join the trade organization. *[Political Economy]*

(2) 那幹嘛不讓萊豬進入校園
Then why not introduce US pork with ractopamine into the campus? *[Food Safety]*

(3) 過的很好 讓我吃到萊豬跟渣米 吸到空污
I live a good life. Let me eat US pork containing ractopamine along with rice dregs and breathe polluted air. *[Both]*

(4) ㄚ反正把萊豬給台豬吃就好辣
Anyway, just feed US pork with ractopamine to Taiwanese pigs. *[Other]*

4 Results

This study not only provides the topic modeling results through the LDA-based word clouds and reports on the predicted results obtained by our fine-tuned BERT model but also demonstrates the evolution of public attitudes toward the US pork issue and analyzes the time series data via DFA.

First and foremost, the word clouds of FS and PE comments created by LDA are shown in Figures 1 and 2 to visualize the lexical semantic information and word frequency via font size. As listed in Table 3, we observe that the keywords extracted by LDA appear to be clustered based on the correlation between a food safety perspective and a politico-economic standpoint. However, we argue that there is no clear-cut between the two sub-issues. Taking a closer look at the overlapping keywords which appear in both word clouds, we observe the politico-economic standpoint seems to be interwoven with the food safety perspective on the US pork issue. More precisely, analyzing the top 25 keywords, we find that the overlapping keywords are mostly associated with PE and that only the phrase 核食 (radiated food) is related to FS; hence, we can derive that FS-related keywords are more representative features than PE-related keywords.

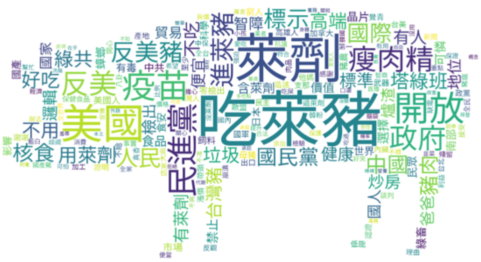


Figure 1. LDA topic model for Food Safety comments.

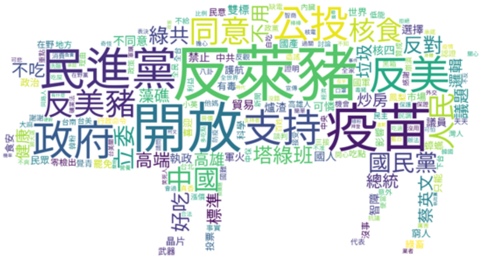


Figure 2. LDA topic model for Political Economy comments.

Category	Keyword
Food Safety	吃萊豬 eat the US pork containing ractopamine, 萊劑 ractopamine, 瘦肉精 a leanness-enhancing feed additive, 不吃 never eat, 好吃 delicious, 用萊劑 use ractopamine, 豬肉 pork, 標準 standard, 標示 label, 美國 USA
Political Economy	反萊豬 oppose the US pork with ractopamine, 反美 oppose the US, 公投 referendum, 支持 support, 同意 agree
Overlapping	反美豬 oppose the US pork, 政府 government, 民進黨 DPP, 國民黨 KMT, 開放 lift a ban, 疫苗 vaccine, 高端 MVC COVID-19 vaccine, 中國 China, 塔綠班 tǎlyùban, 核食 radiated food

Table 3. List of top 25 LDA topic keywords.

Secondly, Table 4 indicates that our fine-tuned BERT model achieves an overall performance of 0.96 in F1-score. A possible explanation for the success of our BERT model in topic classification is the repetitive nature of data. This inference is compatible with Tu et al.'s (2021) finding that key opinion leaders regularly post and repost articles regarding the US pork issue on PTT to stimulate discussion among netizens. Furthermore, Table 5 gives a comparison of the number of comments for each category between the annotated subset and the whole dataset. Since the comments predicted as FS and PE by our BERT model account for more than half of the total comments, we can confirm that the large majority of netizens look at the US pork issue either from a food safety perspective or from a politico-economic point of view.

Category	Precision	Recall	F1-score	Support
FS	0.95	0.97	0.96	824
PE	0.98	0.97	0.98	1058
Both	0.95	0.96	0.96	503
Other	0.95	0.94	0.95	654
Macro avg.	0.96	0.96	0.96	3039

Table 4. BERT classification results.

Category	# Comments	
	Subset	All
Food Safety	4,168	24,018
Political Economy	5,187	20,454
Both	2,584	11,941
Other	3,251	17,627
Total	15,190	74,040

Table 5. Dataset information.

After the fine-tuned BERT model predicts the whole dataset, the predicted results are given to illustrate the evolution of public opinion over time as can be seen in Figure 3. It is worthy of note that the growth in the number of comments is consistent with the corresponding social events and situations. Setting the number of 100 comments as the benchmark value, we observe four periods of time when a heated discussion among netizens took place, as marked in Figure 3. First, the US pork issue went viral between January and February as the administrative orders concerning the ban lift on US pork came into effect on January 1, 2021. Second, considering the appearance of the high-frequency keyword 疫苗 (vaccine), we maintain that since both the COVID-19 pandemic and the import of US pork are associated with public health, the US pork issue was frequently raised and discussed on PTT between May and June due to a surge in the number of COVID-19 cases and a COVID-19 vaccine shortage at that time. Third, given the high-frequency keyword 高端 (MVC COVID-19 vaccine), a sudden increase in late August may relate to the vaccination of the first dose of the MVC COVID-19 vaccine in Taiwan on August 23. Fourth, since the referendum against the US pork importation took place on December 18, 2021, netizens on PTT were engaged in a heated discussion over the US pork issue as the end of 2021 approached.

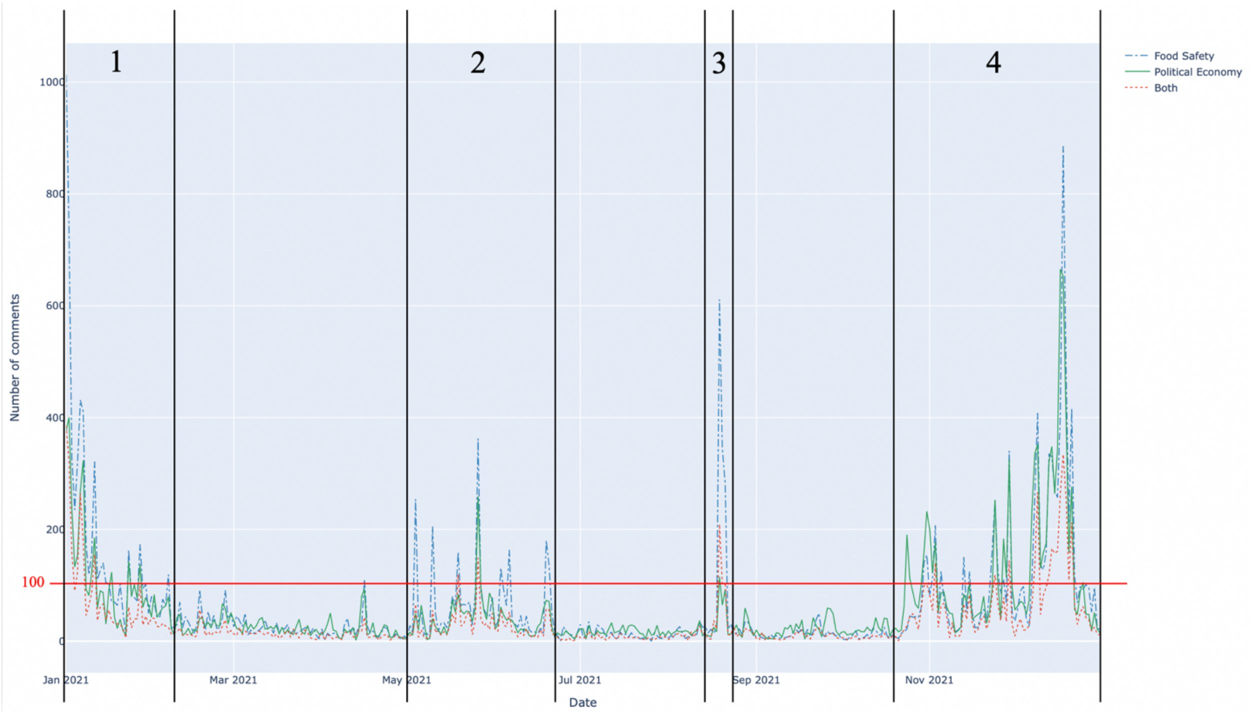


Figure 3. Evolution of public opinion.

As illustrated in Figures 4, 5, and 6, the notable finding based on DFA is a significant rising trend in the number of FS, PE, and Both comments over time in the year 2021 without any downward trend. The estimation of the scaling exponent $\boxed{\propto = 0.67}$, $\boxed{\propto = 0.65}$, and $\boxed{\propto = 0.67}$ for the FS, PE, and Both comments, respectively, demonstrates the long-range correlations in the number of comments per day. Though the overall number of FS comments per day surpasses that of PE comments, the proximity of the exponents obtained by the DFA method infers the absence of a more salient perspective between the FS and PE sub-issues. This finding is in accordance with the fact that the pork import referendum failed to pass but only by a narrow margin. While 51.21 percent of voters rejected the opposition against the import of US pork, there were still 48.79 percent of voters advocating a ban on the US pork importation. The referendum result implies that the potential disadvantage of food safety hazard does not outweigh the politico-economic benefits despite a larger volume of online discussion on the FS sub-issue throughout the year 2021.

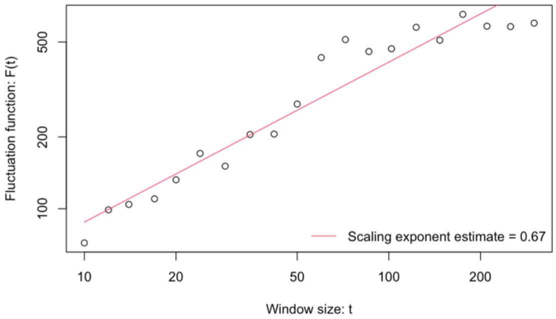


Figure 4. Detrended fluctuation analysis of Food Safety comments.

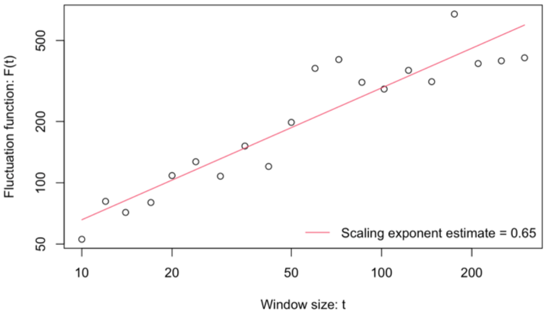


Figure 5. Detrended fluctuation analysis of Political Economy comments.

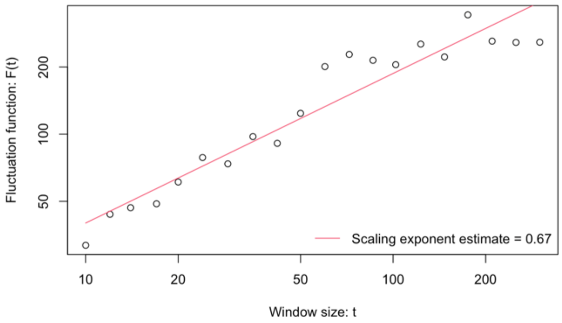


Figure 6. Detrended fluctuation analysis of Both comments.

5 Conclusion

This study investigates how netizens on PTT perceive the US pork importation by examining the interaction between the FS and PE two sub-issues through the combination of machine learning and quantitative analysis. In order to retrieve useful information from a large number of comments on PTT related to the import of US pork, we adopt a complementary approach using both unsupervised and semi-supervised transfer learning. First, at the word level, we incorporate LDA into word clouds to identify the main topics and assess whether the extracted keywords are clustered according to the correlation between the two sub-issues. Second, at the sentence level, we annotate the subset extracted from the whole dataset. Then, we treat the annotated subset as the gold standard to fine-tune a BERT model extended to make predictions on the whole dataset. Finally, at the time series level, we use the DFA to analyze the number of FS and PE comments per day.

Our main results are the following: (1) the finding that the FS sub-issue interweaves with the PE sub-issue in terms of the US pork importation at both word and sentence levels; (2) the absence of a more prominent perspective between the two sub-issues at the time series level; (3) the development of a semi-automated annotation approach and annotation guidelines for a domain-specific topic model; (4) the performance of 95% in F1-score for topic detection achieved by our fine-tuned BERT model; (5) a chronological view of public opinion predicted by the application of our fine-tuned BERT model to the downstream topic classification task; (6) the observation of a steady increase in the number of FS and PE comments over time in the year 2021 based on the DFA.

One potential limitation of this study is the inevitability of prediction error when applying the predicted results obtained by the fine-tuned BERT model to the DFA procedure. Accordingly, we manually label as much data as possible and make our BERT model as accurate as possible during the fine-tuning process in order to minimize errors. Despite possible errors predicted by the model, our proposed approach still offers a data-driven method for quantifying public perceptions of a particular issue and provides new insights into public opinion on the US pork issue based on the integration of machine learning and quantitative analysis.

This paper has contributed to the growing body of studies on social media mining by probing the dynamics of public opinion from a linguistic perspective and examining the interplay between sub-issues of a specific social issue at the word, sentence, and time series levels. It would be useful to conduct more research on the evolution of public attitudes toward different social issues to add further depth, richness, and insights to the findings of this study and provide a further development of topic classification and time series analysis of textual data on social media.

Acknowledgments

We are grateful to four anonymous reviewers for their constructive and insightful comments, which have been taken into account in the revised version to improve the quality of this study. We would also like to thank Dr. Hen-Hsen Huang for his valuable feedback. Finally, this project is supported by a research grant awarded by the National Science and Technology Council, Taiwan (NSTC 111-2410-H-004-098).

References

Anstead, N., & O'Loughlin, B. 2015. Social media analysis and public opinion: The 2010 UK general election. Journal of computer-mediated communication, 20(2), 204–220.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3, 993–1022.

Brulle, R. J., Carmichael, J., & Jenkins, J. C. 2012. Shifting public opinion on climate change: an empirical assessment of factors influencing concern over climate change in the US, 2002–2010. Climatic change, 114(2), 169–188.

Chuang, J. H., & Hsieh, S. K. 2015. An Arguing Lexicon for Stance Classification on Short Text Comments in Chinese. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters (pp. 27–36).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Faulkner, A. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, 174–179.

Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), 1735–1780.

Hong, L., & Davison, B. D. 2010. Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80–88).

Houlsby, N., Giurgiu, A., Jastrzębski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790–2799). PMLR.

Kantelhardt, J. W. 2011. Fractal and multifractal time series. In Meyers, R. A. (Ed.), Mathematics of complexity and dynamical systems (pp. 463–487). Springer.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. 2011. Twitter trending topic classification. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 251–258). IEEE.

Li, W., & McCallum, A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577–584).

Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1–167.

Peng, C. K., Halvin, S., Stanley, H. E., & GoldbergerA, L. 1995. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos: An Interdisciplinary Journal of Nonlinear Science, 5(1), 82–87.

Stimson, J. A. 1999. Public opinion in America; moods, cycles and swings, CO: Westview Press.

Tu, S. T., Wu, C. Y., Lu, L. Y., & Hsieh, C. H. 2021. Opinion Leaders in Internet Public Opinion: A Case Study on the Controversy of US Pork Containing Ractopamine. In 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (pp. 1–2). IEEE.

Valdez, D., Pickett, A. C., & Goodson, P. 2018. Topic modeling: latent semantic analysis for the social sciences. Social Science Quarterly, 99(5), 1665–1679.

Walker, M., Anand, P., Abbott, R., & Grant, R. 2012. Stance classification using dialogic properties of persuasion. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 592–596).

Wang, S. I., & Manning, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 90–94).

Yin, W., Hay, J., & Roth, D. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLPIJCNLP 2019, Hong Kong, China (pp. 3912–3921). Association for Computational Linguistics.

Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons

Akila Peiris and **Nisansa de Silva**
Department of Computer Science & Engineering,
University of Moratuwa, Sri Lanka
{akila.21,nisansadds}@cse.mrt.ac.lk

Abstract

This paper introduces the Forgotten Realms Wiki (*FRW*) data set and domain specific natural language generation using *FRW* along with related analyses. Forgotten Realms is the de-facto default setting of the popular open ended tabletop fantasy role playing game, Dungeons & Dragons. The data set was extracted from the Forgotten Realms Fandom wiki consisting of more than over 45,200 articles. The *FRW* data set is constituted of 11 sub-data sets in a number of formats: raw plain text, plain text annotated by article title, directed link graphs, wiki info-boxes annotated by the wiki article title, Poincaré embedding of first link graph, multiple Word2Vec and Doc2Vec models of the corpus. This is the first data set of this size for the Dungeons & Dragons domain. We then present a pairwise similarity comparison benchmark which utilizes similarity measures. In addition, we perform D&D domain specific natural language generation using the corpus and evaluate the named entity classification with respect to the lore of Forgotten Realms.

1 Introduction

Specialized and domain specific data sets are useful for a number of advanced tasks in the domain of Natural Language Processing (NLP). For example, recent studies have shown that the domain specificity significantly impacts vital NLP tasks such as measuring semantic similarity (Sugathadasa et al., 2017) and domain specific text generation (Lebret et al., 2016). Further, it has been shown that models developed using data from a generic domain do not seamlessly transfer to tasks in a specific domain (Rajapaksha et al., 2020). Fantasy domains could be considered an extreme case of domain specific data, as it is possible to observe the full spectrum of deviations from the non-domain specific (general domain) usage, both in the lexical and semantic perspectives. An example for lexical differences is the usage of *dwarves* as the plural form of *dwarf* in the fantasy genre[1] in place of the general domain spelling *dwarfs*. An example for semantic differences can be seen in the words *Ghost*[2] and *Wraith*[3]. In the Merriam-Webster dictionary, they are given as synonyms in the generic domain[4] while in the domain of the fantasy role playing game Dungeons & Dragons, they are defined as two distinct creatures. In this paper, we present the *FRW* data set, specific to the *Forgotten Realms* setting of *Dungeons & Dragons*. We expect our data set to be useful for in-domain tasks such as text generation (Zhang et al., 2019), information extraction (de Silva and Dou, 2021), and information retrieval (Sugathadasa et al., 2018). We also anticipate our data set being vital for cross domain tasks such as text alignment (Sanchez-Perez et al., 2014), style transfer (Fu et al., 2018), and summarizing (El-Kassas et al., 2020). As a primer for these usages, we introduce a pairwise similarity comparison benchmark and evaluate the domain-specific free text generation task.

1.1 Dungeons & Dragons

Dungeons & Dragons (D&D or DnD), is an open-ended pen and paper tabletop role playing game (RPG) which has been commercially available since Gygax and Arneson (1974) published the first version. The games are primarily based on fantasy genre. However, there is a plethora of other settings ranging from science fiction, post-apocalyptic to hollow world and much more. Even within a selected genre, it is highly customizable, for example, a fantasy setting might be in high or low fantasy. D&D has a predefined set of rules governing almost every aspect of the gameplay including the *setting*. A setting has a lore, species and artifacts among other components; which can be dissimilar

[1] This is inherited from the spelling used in the *Lord of the Rings* and other relevant publications by J. R. R. Tolkien

[2] https://bit.ly/DnDGhost

[3] https://bit.ly/DnDwraith

[4] https://bit.ly/3Z2YsHC

between settings. There are also several editions of D&D, with 5 (Crawford et al., 2014) being the latest. It is the version that our *FRW* data sets are predominantly based on. However, it does contain some information from earlier editions in cases where there have been changes to the lore between versions or in cases where information have been consistently brought forward.

1.2 Forgotten Realms Wiki

Forgotten Realms as mentioned, is a setting which is categorized under *high fantasy*, set in an alternate world filled with magical elements combined with larger than life themes, plots, and characters. It originated as a medieval European setting but over the years, has been influenced by other cultures including Middle Eastern and Asian. *Forgotten Realms* became the most utilized of all the official D&D settings after it became the de-facto default setting of the immensely popular (Whitten, 2021) 5th edition. Almost all of the official material published for D&D is based on this setting. Due to this, *Forgotten Realms* now has the most resources and information available from all the settings in D&D.

However, this massive amount of information is distributed among hundreds of official books and magazines making it intractable for a casual enthusiast of D&D. To remedy this problem and to curate and consolidate the information, the community of D&D enthusiasts voluntarily contribute and maintain the *Forgotten Realms wikia* [5]. A *Wikia* or a *Fandom Wiki* is a Wikipedia [6]-esque website (uses the same MediaWiki [7] collaborative documentation platform) hosted by Fandom, Inc. [8]. This is typically dedicated to a particular domain (e.g., Star Wars [9], Marvel [10], Harry Potter [11], Formula One Racing [12]). The *Forgotten Realms Wikia* has over 45,200 articlesas of September 2022 and keeps growing at a rapid pace.

1.3 Wikipedia and other Wikis as Data Sources

Wikipedia and other Wikia, maintained by a community of volunteers, are treasure troves of domain specific knowledge (Ferrari et al., 2017). While there are endless debates regarding the validity of such community maintained knowledge-bases in scientific context (Cozza et al., 2016; Ferschke, 2014), there are still a number of ways they can be used to further the scientific frontiers (Ponzetto and Strube, 2007; Zesch and Gurevych, 2007; Zesch et al., 2008). One such usage in the field of Natural Language Processing is to use them as data sources which not only provide corpora of the relevant domains but also provides insight into community-based collaborative text maintenance (Ferschke et al., 2013).

The possibility of accessing as a freely available source in multiple languages (Nastase and Strube, 2013) (human translated), being extensive, and having special information content such as infoboxes[13] make Wikipedia and similar wikia rich resources for data. An infobox is the table-like structure typically found at the top-right side of a wiki article. It is a human annotated, tabular summary of the article, arranged in a key-value structure according to a template. According to Lange et al. (2010) about one third of all Wikipedia articles contain an infobox. While this is indeed a rich source of information, they are known to be noisy and sparse(Hoffmann et al., 2010). The wiki page itself only renders the pairs that contain values.

Another special information content is found in the first paragraph/ (lead section [14]) of a wiki article. According to the guideline, this is typically formatted as an abstractive summary to the entire page. In their study on wikipedia, Lange et al. (2010) report that there is a 92% chance to find any of the information summarized in the infobox within the first paragraph.

1.4 Domain Specific Text Generation

Domain specific text generation is an emerging area in NLP (Liu et al., 2018; Chen et al., 2021; Zhang et al., 2022; Amin-Nejad et al., 2020). The objective in this is to generate text which adheres to a given domain, in the sense that the content generated should be semantically and pragmatically truthful to the said domain. One of the reasons why domain specific text generation is difficult compared to generic text generation is that, in most cases this requires copious amounts of linguistic resources based on the domain in question. This hurdle is true even for fine-tuning a pre-trained

[5] https://forgottenrealms.fandom.com

[6] https://en.wikipedia.org

[7] https://bit.ly/3YHpG6K

[8] https://www.fandom.com

[9] https://starwars.fandom.com/

[10] https://marvel.fandom.com/

[11] https://harrypotter.fandom.com/

[12] https://f1.fandom.com/wiki/

[13] https://bit.ly/3lLcqiOen.wikipedia.org/wiki/Help:Infobox

[14] https://bit.ly/3IAzHNx

model which relatively demands less amount of data than training a model from ground-up (Zhang et al., 2021).

2 Related Work

2.1 Wikipedia and Other Wiki Data Sets

In recent times the availability of linguistic data sources have increased significantly. Especially Wikipedia based data sets such as Wit (Srinivasan et al., 2021), WCEP (Ghalandari et al., 2020), and SQuAD (Rajpurkar et al., 2016). Tools such as LUCHS (Hoffmann et al., 2010) and WOE (Wu and Weld, 2010) are capable of extracting information from Wikipedia pages to create such data sets. Both systems rely on the key-value structure of the infoboxes to guide the information extraction process from the natural language text. This guided process is akin to the widely used Ontology-Based Information Extraction (OBIE) (de Silva et al., 2017).

As mentioned, Fandom, Inc. is an organization which hosts wikis for a large number of entertainment media franchises and other areas as the general populace may desire. The Fandom wikis operate on the same technology and guidelines[15] as Wikipedia. They are good sources of domain specific data for different media franchises as they are written in the desired domain and offers a clear demarcation from in-domain and out-domain data as opposed to obtaining data from sources such as the common crawl (Kreutzer et al., 2022). The Critical Role Dungeons & Dragons Data set (*CRD3*) (Rameshkumar and Bailey, 2020) is a D&D domain-specific, narrative driven, multi speaker dialog data set that has been extracted from a similar Fandom Wiki[16]. This particular wiki is dedicated to the web series, *Critical Role*, a live D&D gameplay series. The data set consists of multi-speaker dialogue that form a narrative, paired with their abstractive summaries.

2.2 Domain Specific Text Generation

Text generation methodologies fall into three categories (Stent et al., 2004). **Template based** methods (Busemann and Horacek, 1998; Reiter and Dale, 1997; McRoy et al., 2003) are the most common variant. It uses pre-defined text templates applicable to different scenarios to generate text. It is a tedious and non-salable approach. Secondly, there is **Rule based** generation (Dale et al., 2003; Turner et al., 2009; Reiter et al., 2005). This has three inter-dependent phases: (1) text planning - governs the process of meaning representation retrieval from a knowledge base, (2) sentence planning - governs the words and their order to produce coherent sentences, and (3) surface realization - converts the sentence plan into actual sentences. Thirdly, the **Data driven** approach (Barzilay and Lapata, 2005; Liang et al., 2009). Unlike rule based approaches, data driven ones require more data. This burden is alleviated using pre-trained language models and transfer learning techniques. Open AI's Generative Pre-trained Transformer models GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) as well as the open source reproductions of these, the GPT-Neo models (Gao et al., 2020; Black et al., 2022) are such large pre-trained language models. In addition to having been trained on very large data sets, they are also large networks. These models are capable of generating highly sophisticated texts. With some fine-tuning, they can be adapted to do the same for specific domains.

3 Forgotten Realms Wiki (*FRW*) Data set

We introduce the Forgotten Realms Wiki (*FRW*) data set[17], extracted from the Forgotten Realms Fandom wikia. We have extracted multiple data sets from this textual resource and present them in multiple formats for different linguistic use cases. We also present several different embeddings for this data set including Poincaré hierarchical embedding and multiple word and document level embeddings. A summary of the data sets is shown in Table 1 and the individual statistics for each data set can be found listed under Table 2.

The plain text corpora (*FRW-P*, *FRW-J*) are devoid of special data structures and other markings. As for the links, the MediaWiki allows having an alternative text to display for the links instead of the exact page title for aesthetics of the writing, hence we extract that part for the plain text corpus when available. The *FRW-FJ* data set is composed of mainly the *lead sections*. Because of this, we can consider this as an abstractive summary of *FRW-J*. The *FRW-CL* links pages with categories. The categories themselves have rendered pages which aggregate the pages under each category. The infobox data in *FRW-I* are converted from markdown to

[15] `https://www.mediawiki.org/`
[16] `https://criticalrole.fandom.com`
[17] `https://huggingface.co/datasets/Akila/ForgottenRealmsWikiDataset`

JSON before being embedded in the overall JSON structure indexed by the page title. Each of the word and document embedding data sets (*FRW-W*, *FRW-D*, *FRW-FD*) have 2 different embeddings used. For word level embeddings CBOW and Skip-gram (Mikolov et al., 2013) are used. While PV-DBOW and PV-DM (Le and Mikolov, 2014) are used in document embeddings. Figure 1 shows the convergence of the Poincaré embedding data set, *FRW-PE*.

All of the data sets that use a JSON structure (*FRW-J*, *FRW-FJ*, *FRW-I*) use the same high level JSON schema. The pages are organized in a JSON array with *page* and *content* being the only two attributes in each element. The *page* attribute contains the article title while the *content* attribute contains the corresponding information extracted from the page. This information is in plain text format for *FRW-J* and *FRW-FJ*. In the case of *FRW-I*, it is a JSON dictionary containing the infobox content as the key-value pairs. Code 1 shows the top level JSON schema.

Code 1: JSON top level structure

```
[
  {
      "page": "page_title",
      "content": "page_content"
  },
  ...
]
```

Name	Description
FRW-P	Raw plain text corpus (no Markdown text markings)
FRW-J	A JSON structure with plain text indexed by article title
FRW-FJ	A JSON structure with only the first paragraph (plain text) of the articles indexed by article title
FRW-L	A directed graph indicating all the references in the articles to other articles
FRW-FL	A directed graph indicating the first references in the articles to other articles
FRW-CL	A directed graph indicating the category references in the articles to category pages.
FRW-I	A JSON structure for the Wikipedia infobox substructures indexed by article title
FRW-PE	Poincaré embedding of *FRW-FL*
FRW-W	2 Word2Vec models for *FRW-P* (CBOW and Skip-gram)
FRW-D	2 Doc2Vec models for *FRW-P* (PV-DBOW and PV-DM)
FRW-FD	2 Doc2Vec models for *FRW-FJ* (PV-DBOW and PV-DM)

Table 1: *FRW* data set

4 Use Case Analysis 1: Semantic Similarity Comparison

To illustrate both the consistency as well as the non-trivial nature of data sets we have collected, we have performed similarity calculations for a set of text pairs extracted from the data set using multiple different similarity metrics. By the high alignment of semantic similarity in similar perspective data sets, we show the consistency in our data sets. By the low alignment of the differing perspective data sets, we show that the individual data sets are not redundant and that they carry unique information that may not have overlaps with other data sets that we present in this work.

4.1 Text Pairs for Evaluation

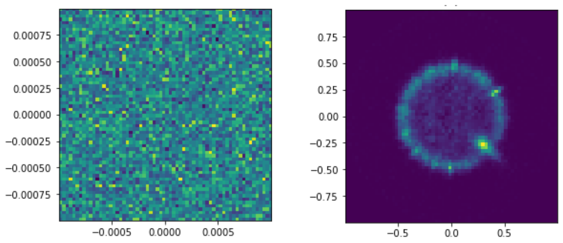
To ensure that these different metrics are comparable, we have used the same set of text pairs for all of the similarity calculations. Hierarchical similarities are measured using article titles and the embedded vector distance based similarities are calculated using article contents. We use the *FRW-FJ* data set to generate the text pairs. Since the *FRW-FJ* data set is a subset of *FRW-J*, it ensures that a) all the pairs correspond to actual wiki articles b) has valid text content c) full document and first paragraph only data sets are available for different similarity calculations, and d) almost all nodes (page titles) are significant/“worthy of notice”[18] to the domain as per the Wikipedia guidelines.

4.1.1 First Link

First link is the first internal reference link (refers to another article in the same wikia) found in an article that is not a broken link or a miscellaneous link such as the pronunciation guide. According to the wikimedia guidelines, the lead section of a typical Wikipedia article contains links to other articles that provide context to the article in question i.e., the references in lead section point towards more generalized concepts and/or any other concepts related to the context of that article. We use this arrangement to measure the similarity or relatedness of topics. This leads to an interesting pattern where clicking the the first link of a random Wikipedia page and doing so repeatedly on the subsequent pages will 97% of the time (Lamprecht et al., 2016) lead to a cycle containing the article “Philosophy”[19] . The rest of these *first link*

[18] https://bit.ly/3S9HHbc

[19] https://bit.ly/3xyBnRf



(a) Initial (0 epochs)

(b) After 50 epochs

Figure 1: Poincaré Embedding convergence

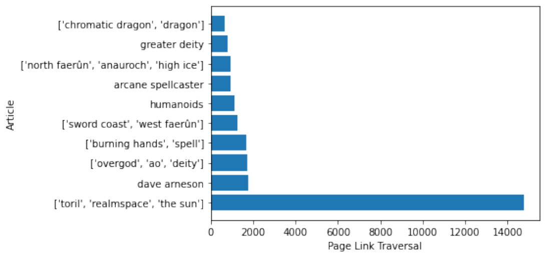


Figure 2: First link traversal graph

Statistic	Value
Total number of tokens (excluding titles)	9,189,536
Total number of tokens (including titles)	9,287,670
Total number of unique tokens	145,624
Total number of sentences	517,248

(a) *FRW-P*

Statistic	Value
Total number of articles	48,892
Average number of tokens per sentence	17.77
Average number of tokens per article	187.96
Average number of sentences per article	10.58

(b) *FRW-J*

Statistic	Value
Total number of articles	41,204
Total number of tokens	980,047
Total number of sentences	98,244
Average number of tokens per sentence	9.98
Average number of tokens per article	23.78
Average number of sentences per article	2.38

(c) *FRW-FJ*

Statistic	Value
Average number of attributes per infobox	40.54
Average number of completed (filled) attributes per infobox	10.40
Total number of articles containing infoboxes	35,923

(d) *FRW-I*

Statistic	Value
Total number of nodes	46,910
Total number of edges	570,857
Average number of edges per node	12.16

(e) *FRW-L*

Statistic	Value
Total number of nodes	43,329
Total number of edges	41,213
Number of nodes not referenced by others	34,881
Number of nodes with no references	2151

(f) *FRW-FL*

Table 2: Statistics of different sub data sets of the $FRWdataset$

traversals exhibit one of the following shortfalls: 1) contain no internal links, 2) contains a self loop, 3) ends up in an isolated tree, 4) form a cycle with a few other pages. The Forgotten Realms wiki also abide by the same principle. The *center* of the Forgotten Realms wiki universe is a cycle composed of the articles, "*Toril*", "*Realmspace*" and "*The Sun*". However, unlike in the case of Wikipedia, in Forgotten Realms wiki, this only applies to around 30.2% of the articles. Figure 2 lists the 10 most commonly traversed articles using this method. The ones enclosed in "[ ]" refer to cycles, for example ['*toril*', '*realmsphere*', '*the sun*'] refers to a first link cycle between the three corresponding articles.

4.1.2 Issues with Category Links for Semantic Similarity Evaluation

Even though it is the de-facto categorical hierarchy, there are many issues with using category links as a measure for semantic similarity. The most prominent bottleneck of *FRW-CL* data set is that it is mostly a flat hierarchy. So any set of node pairs would have almost identical distance measures no matter how different they are semantically. Secondly, the Categories are not consistent across all articles, i.e,. while some articles may have an abundance of Categories, others may have have little to none. Finally, Category pages do not necessarily have article content as a typical page does, hindering the ability to perform effective word and document embedding.

4.1.3 Generating Text Pairs for Evaluation

We created 1,000,000 unique text pairs using 41,000 nodes from *FRW-FL*. We have also ensured that there are no interchangeable duplicates. To ensure that the selected pairs have better representation, we have used a weighted random sampling technique with dynamically updated weights. The sampling was done with replacement. The probability of an item i getting selected for the sample pair set is given in Equation 1, where N is the total number of pairs we generate and k_j is the number of times the jth item has already been selected.

4.2 Hierarchical Similarity Measures

We have used the *FRW-FL* data set as the base for similarity measures using hierarchical similarity evaluation methods. Although the *FRW-FL* is already devoid of any self-loops, there are cycles and isolated trees while also lacking a common root node. We process this and convert into a directed cyclic graph.

Let G be a graph in the set of disconnected graphs $G = (V, E) \in G_1, G_2, ..., G_n$ where $E \in E'$ and $V \in V'$. G_c represent a subgraph of a given graph $G_c = (V_c, E_c) \in G$ where $e_1, ..., e_n$ is a trail with vertex sequence $a_1, ..., a_n$ (cyclic graph). Then $\forall G \in G_1, ..., G_n$ apply Equation 2 to obtain the final unified graph, $G' = (V', E')$.

For the intermediate node v', we use a comma separated combination of the names of the nodes forming the cycle. Using an intermediate node helps us retain the relatedness they had with one another up to a certain degree before reaching the root node. We use algorithmic measures such as Wu and Palmer (1994) similarity metric and Jiang and Conrath (1997) distance measure, both of which use the Least Common Ancestor (LCA) as the basis for the calculations. Apart from this, we have also evaluated with the hierarchical embedding using Poincaré (Nickel and Kiela, 2017) method. While other embedding methods such as ones using metadata (Xing and Paul, 2017; Zhou et al., 2015) can be experimented with, we have chosen the Poincaré embedding since we are measuring hierarchical similarities. This allows us increase the comparability between the different types of measurements in our experiment.

It should be noted that, for the sake of this comparative analysis, we have converted the Jiang-Conrath distance measure (Jiang and Conrath, 1997) into a similarity measure ranging from 0 to 1 as shown in the Equation 3 where the $LCA(a, b)$ function returns the least common ancestor of the nodes a and b, the $IC(d)$ function returns the Information Content of the node d, and c_i is the node in the hierarchy representing the term t_i.

4.3 Embedding Based Similarity Measures

We have performed both word embedding on *FRW-P* corpus and document embedding for the *FRW-J* data set to create *FRW-W* and *FRW-D* data sets. Both of these data sets contain essentially the same content albeit the format. In addition, we have created *FRW-FD* data set using *FRW-FJ* which only contains the first paragraph of each page to evaluate the effectiveness of the first paragraph in comparison to the whole text. For embedding vector distance based similarity, we have used the *FRW-W* data set containing CBOW and Skip-gram (Mikolov et al., 2013) model embeddings. For word embedding, when a title is given the corresponding article is retrieved from *FRW-J*. Then for all the words in the article, the word vectors are fetched from the *FRW-W* data set and a single vector is obtained via average pooling. Cosine similarity is defined as shown in Equation 4, where t_i is a string in the domain and v_i is the corresponding vector in the embedded vector space.

$$P(i) = \frac{\sqrt{N} - k_i}{N\sqrt{N} - \sum_{j=1}^{N} k_j} \tag{1}$$

$$E' = \begin{cases} E \cup (root, v) & \text{if } deg^{-}(v) = 0 \text{ and } v \neq root \\ E \cup (root, v') \cup (v', v) & \text{if } v \in G_c \\ E & \text{otherwise} \end{cases} \tag{2}$$

$$JC\_s(t_1, t_2) = \frac{1}{1 + |2 \times IC(LCA(c_1, c_2)) - (IC(c_1) + IC(c_2))|} \tag{3}$$

$$cosine\_similarity(t_1, t_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||} \tag{4}$$

For Doc2vec (Le and Mikolov, 2014), we have used both *FRW-D* and *FRW-FD* data sets with each having a `PV-DBOW` and a `PV-DM` model (Le and Mikolov, 2014) resulting in four embedded models altogether. The four models are trained with the article title as the tag for the content of the article. Hence the document vector itself can be fetched directly from the model using the article title (text phrase).

We have briefly mentioned at the start of this section, we specifically used the first paragraph only text to assert for its goodness compared to the whole text. The rationalé for this as follows: as discussed in subsection 4.1.1, the first paragraph or the lead section of a wiki article is an abstractive summary of the entire article. Hence, if this showed comparable results to full text, the full text document embedding can be substituted by this. Which requires less computational resources due to its much smaller size. For comparison, the word count on data sets *FRW-J* and *FRW-FJ* are 9,189,536 and 980,047 respectively, which is a 10:1 compression ratio. Further, this would open the door to future in-domain text summarizing research.

				Hierarchical			Embedding					
							Word2Vec		Doc2Vec			
				WP	JC	P	(*FRW-J*)		(*FRW-FJ*)		(*FRW-J*)	
							CBOW	SG	DM	DBOW	DM	DBOW
Hierarchical			WP	1.0000								
			JC	0.6346	1.0000							
			P	0.0097	0.0624	1.0000						
Embedding	Word2Vec	(*FRW-J*)	CBOW	0.0581	0.0212	0.0013	1.0000					
			SG	0.0553	0.0188	-0.0043	0.9412	1.0000				
	Doc2Vec	(*FRW-FJ*)	DM	0.0040	-0.0298	0.0548	-0.0626	-0.0791	1.0000			
			DBOW	0.0466	0.0155	0.0359	0.0362	0.0222	0.5691	1.0000		
		(*FRW-J*)	DM	0.0259	0.0186	0.0175	-0.1865	-0.2593	0.1724	0.1484	1.0000	
			DBOW	0.0361	0.0287	0.0453	-0.0896	-0.1601	0.1511	0.1825	0.5493	1.0000

Table 3: Pearson Correlation for the pairwise text similarities across multiple similarity metrics: 1) WP – Wu & Palmer similarity 2) JC – Jiang-Conrath Similarity 3) P – Poincaré metric 4) CBOW – Continuous Bag of Words 5) SG – Skip-gram 6) DM – Distributed Memory 7) DBOW – Distributed Bag of Words

5 Use Case Analysis 2: Forgotten Realms Free Text Generator

Our work on free text generation[20] is divided into two phases. In the first phase, we try to train a GPT-Neo (Gao et al., 2020) model capable of generating text in the Dungeons & Dragons (Forgotten Realms) domain. Once the text generator was in place, the second phase of the work involves classifying in-domain named entities present in the generated text. We used the trained model from step 1 and fine-tuned it using data from the Forgotten Realms Wiki JSON (*FRW-J*) database. We tokenized each of the the documents in the *FRW-J* data set separately. The model that was used contained 1.3 billion parameters. Using this fine-tuned model, we were able to generate texts that are grounded in the D&D domain.

5.1 Generated Text

An example text generated using the Forgotten Realms Free Text Generator we have trained using the *FRW* data set is shown in Figure 3. To make comparisons as to the effect of training on the *FRW* data set (fine-tuning, we have also listed some text samples generated using the unmodified (vanilla) version of the same language model (Figure 4). All of these samples were generated based on the same input prompt "*A Bard, a Wizard, and a Paladin walk into a tavern*".

We can observe, that the vanilla model refers to real world entities such as *Northumberland*, *Emerald Tablet*, and *Alderney*. It is also worth noting that the vanilla model refers to other works of literature as can be observed in Figure 4a where *Order of the Phoenix*[21] was mentioned. Here, the entity was correctly identified as a group or a society in accordance with its literature and has been used appropriately in the generated text.

In comparison, the text samples generated from

A Bard, a Wizard, and a Paladin walk into a tavern, and debrief the group on their travels. The trio discuss their lives and lives. Bard: "In actuality, the only things I have gained in my life that I truly value are knowledge and comradeship, both of which seem more valuable than wealth." Wizard: "Clan Bloodforge was once my clan; I guess I'm considering it my clan after death to boot.". Paladin: "There are two separate communities of people living in this land, one in the North and another in the South, and the northern people came to the west to the Sword Coast about five centuries ago." One of Bloodforge's leaders meets with the leaders of Clan K'yorlhan, the tribe from Tethyr that fights the bard raider bandit leader Korgh, and discovers that most of the members of Clan Bloodforge do not want to join Clan K'yorlhan. In the end, he leads them to the hidden Bloodforge, and joins them in battle. Monk: "Korgh is a madman, but he's got a point about the invaders.". Bard: "That's true; otherwise, why'd he attack all across Tethyr? That's something Clan Bloodforge are going to have to consider." During the homecoming party, Bard, Wizard, and Paladin trade news of their adventures with each other. Bard: "Clan Bloodforge and Clan K'yorlhan, now that's something I'm going to have to watch carefully.". Wizard: "I'd better keep in mind that Clan K'yorlhan is already taken.". Paladin: "We were also allied with the tanarukks of Norland, but I hope I'm wrong, because they keep countless spies hidden around here.". Paladin - Son of Blackstaff - joins Bloodforge as a young man. Bard: "Artemis Entreri is not the most forward-thinking of consorts, but still, he's a good man." Paladin - A young warrior from the north. Bard: "Welcome to Bloodforge." Paladin - A young man from the west. Bard: "The Wyrmskull Thron is a noble house in Calimshan.". Wizard: "I'd have to say that the War Wizard of Calimshan is a good man."

Figure 3: Sample text generated from the Forgotten Realms Free Text Generator (trained on *FRW* data set)

[20] https://huggingface.co/Akila/ForgottenRealmsFreeTextGenerator

[21] https://bit.ly/3Z4vOFI

A Bard, a Wizard, and a Paladin walk into a tavern, It was the middle of the night, and - like the rest of the town - Bard, Wizard, and Paladin had already found a place that they could call their own. It was a place where they could help young minds and provide a positive role model for young people. "Do you mind if I sit?" a BaldDwarf asked. "The kids love your games," the Bard replied. The Bard was an active member of the Order of the Phoenix, the Order of the Emerald Tablet, an Amity Group, the Viscount of Ce's Guild, and anointed as a Paladin by a bard named D'Cel. He was also a Wilderness Shield Master and a beardless man. He couldn't wear any clothing he didn't own, because - to him - the Wilds were his home. The Wizard and Paladin should never have come into a tavern, but the Bard was a musketeer, and a Tourney-member. He was just trying to get a look at the other two.

(a) Sample 1

A Bard, a Wizard, and a Paladin walk into a tavern, in rural Northumberland. The Bard and the Wizard pay and are received in the company of a Salmon, and the Paladin is greeted as a traveling merchant by merchants and peasants alike. The Bard is either still in his role as an entertainer, or has passed into the service of some people who know what they are doing. The Wizard has given up his peddling. He is now performing services as an entertainer, singing and dancing and acting as peacemaker or mediator. The Paladin has come here to seek the largess of the Bard, to borrow from him with interest, and to bestow upon him the reward of his protection and the services of his wizardry. The Bard has accepted his terms, and has given his lodgings - a mansion built in a ruined mill, called Cinann House, on the banks of the River Barrow, once belonging to the Earl of Northumberland, who was the brother of stood as guardian of the booth of Alderney who came to trade with Scandinavia to the east.

(b) Sample 2

Figure 4: Sample text generated from the vanilla text generator

the Forgotten Realms Free Text Generator show more D&D domain specific characteristics. It uses established entities from the D&D lore such as *Bloodforge*, *Norland*, and *Calimshan*. It also identifies and uses *Norland* as a location which is part of *Sword Coast* in accordance with the domain data. Another thing to note is that the model even generates fake names and characters that are not mentioned in the data set such as *Korgh* and *K'yorlhan* that fit in well with the fantasy genre and build narratives around those characters. Despite, some minor issues with cohesion, overall, it generates satisfactory results.

5.2 Named entity classifier

Although the Forgotten Realms Free Text Generator managed to create text based on D&D domain, when observed carefully by domain experts, there were some inconsistencies with the established lore of the domain. For example, according to the Forgotten Realms lore, *Artemis Entreri* is an *assassin* and not a *consort* while the *Wyrmskull Throne* is a physical object, not the name of a house. To assess the categorical validity of the named entities generated in the text, we have trained the same model on a data set where each row contains a full text generated by the Forgotten Realms Free Text Generator, a named entity in that text, and the matching category extracted from the *FRW-I*. By performing 5-fold cross validation, we were able to train our model to identify the category for a named entity in a generated text. For this basic analysis, we created 100 instances each containing on average 351.4 words and 19.07 sentences. The model was capable of predicting the correct category with 99.3% accuracy on average, attesting to the power of GPT-Neo (Gao et al., 2020) as well as the potential in domain specific text generation. Since the correct classifications are a set of rules declared by the *FRW-I* data set, and the GPT-Neo model uses a data driven training approach, this can be the first step towards creating a conditional text generator that will bridge the traditional rule-based text generation methods and the novel data-driven methods.

As for the vanilla model, we were unable to perform any meaningful entity classification in relation to the D&D domain, as there were no D&D specific entities that were mentioned in the generated text.

6 Conclusion and Future Work

When performing domain specific text generation, it is important that the output stays true to source material. For this, sufficient data from the domain is required. Other than the raw corpora, additional supplementary data structures such as tabular summaries can help ease the process of evaluating the consistency of generated text in context to the domain. In this paper we present a data set based on the D&D domain and a system that is capable of generating free text that stays consistent to the domain. Apart from this, the named entity classifier model shows promising results as part of a guided text generation system. We hope that the *FRW* offers a convenient unique data set for the D&D domain. We hope that the data set can also be enhanced in the future including an improved linked list to measure evaluation.

References

Amin-Nejad, A., Ive, J., et al. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th LREC*, pages 4699–4708.

Barzilay, R. and Lapata, M. (2005). Collective content selection for concept-to-text generation. In *Proceedings of the conference on Human Language Tech-*

nology and Empirical Methods in Natural Language Processing, pages 331–338.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Brown, T., Mann, B., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Busemann, S. and Horacek, H. (1998). A flexible shallow approach to text generation. *arXiv preprint cs/9812018*.

Chen, H., Takamura, H., and Nakayama, H. (2021). Scixgen: A scientific paper dataset for context-aware text generation. *arXiv preprint arXiv:2110.10774*.

Cozza, V., Petrocchi, M., and Spognardi, A. (2016). A matter of words: NLP for quality evaluation of Wikipedia medical articles. In *International Conference on Web Engineering*, pages 448–456. Springer.

Crawford, J., Wyatt, J., Schwalb, R. J., and Cordell, B. R. (2014). *Player's handbook*. Wizards of the Coast LLC.

Dale, R., Geldof, S., and Prost, J.-P. (2003). Coral: Using natural language generation for navigational assistance. In *Proceedings of the 26th Australasian computer science conference-Volume 16*, pages 35–44.

de Silva, N. and Dou, D. (2021). Semantic oppositeness assisted deep contextual modeling for automatic rumor detection in social networks. In *Proceedings of the 16th Conference of the EACL: Main Volume*, pages 405–415, Online. ACL.

de Silva, N., Dou, D., and Huang, J. (2017). Discovering Inconsistencies in PubMed Abstracts Through Ontology-Based Information Extraction. In *Proceedings of the 8th ACM BCB*, pages 362–371.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2020). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, page 113679.

Ferrari, A., Donati, B., and Gnesi, S. (2017). Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embeddings. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399. IEEE.

Ferschke, O. (2014). The quality of content in open online collaboration platforms: Approaches to nlp-supported information quality management in wikipedia.

Ferschke, O., Daxenberger, J., and Gurevych, I. (2013). A survey of nlp methods and resources for analyzing the collaborative writing process in wikipedia. In *The People's Web Meets NLP*, pages 121–160. Springer.

Fu, Z., Tan, X., Peng, N., et al. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference*, volume 32.

Gao, L., Biderman, S., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ghalandari, D. G., Hokamp, C., et al. (2020). A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Gygax, G. and Arneson, D. (1974). *dungeons & dragons*, volume 19. Tactical Studies Rules Lake Geneva, WI.

Hoffmann, R., Zhang, C., and Weld, D. S. (2010). Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 286–295.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Kreutzer, J., Caswell, I., et al. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *TACL*, 10:50–72.

Lamprecht, D., Dimitrov, D., Helic, D., and Strohmaier, M. (2016). Evaluating and improving navigability of wikipedia: a comparative study of eight language editions. In *Proceedings of the 12th international symposium on open collaboration*, pages 1–10.

Lange, D., Böhm, C., and Naumann, F. (2010). Extracting structured information from wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1661–1664.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Lebret, R., Grangier, D., and Auli, M. (2016). Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Liang, P., Jordan, M. I., and Klein, D. (2009). Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 91–99.

Liu, T., Wang, K., Sha, L., et al. (2018). Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference*.

McRoy, S. W., Channarukul, S., and Ali, S. S. (2003). An augmented template-based approach to text realization. *Natural Language Engineering*, 9(4):381–420.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nastase, V. and Strube, M. (2013). Transforming wikipedia into a large scale multilingual concept network. *AI*, 194:62–85.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In Guyon, I., Luxburg, U. V., Bengio, S., et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

Radford, A., Wu, J., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rajapaksha, I., Mudalige, C. R., et al. (2020). Rule-Based Approach for Party-Based Sentiment Analysis in Legal Opinion Texts. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 284–285. IEEE.

Rajpurkar, P., Zhang, J., et al. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 5121–5134, Online. ACL.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Reiter, E., Sripada, S., Hunter, J., et al. (2005). Choosing words in computer-generated weather forecasts. *AI*, 167(1-2):137–169.

Sanchez-Perez, M. A., Sidorov, G., and Gelbukh, A. F. (2014). A winning approach to text alignment for text reuse detection at pan 2014. In *CLEF (Working Notes)*, pages 1004–1011.

Srinivasan, K., Raman, K., et al. (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Stent, A., Prasad, R., and Walker, M. (2004). Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the ACL)*, pages 79–86.

Sugathadasa, K., Ayesha, B., et al. (2017). Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6. IEEE.

Sugathadasa, K., Ayesha, B., et al. (2018). Legal Document Retrieval using Document Vector Embeddings and Deep Learning. In *Science and information conference*, pages 160–175. Springer.

Turner, R., Sripada, S., and Reiter, E. (2009). Generating approximate geographic descriptions. In *Empirical methods in natural language generation*, pages 121–140. Springer.

Whitten, S. (2021). Dungeons & Dragons had its biggest year ever as Covid forced the game off tables and onto the web.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the ACL*, pages 118–127.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Xing, L. and Paul, M. J. (2017). Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49, Copenhagen, Denmark. ACL.

Zesch, T. and Gurevych, I. (2007). Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 1–8.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth LREC*.

Zhang, H., Song, H., Li, S., et al. (2022). A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Zhang, T., Kishore, V., Wu, F., et al. (2019). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, X., Jiang, Y., Shang, Y., et al. (2021). DSGPT: Domain-Specific Generative Pre-Training of Transformers for Text Generation in E-commerce Title and Review Summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2146–2150.

Zhou, H., Chen, L., et al. (2015). Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP (Volume 1: Long Papers)*, pages 430–440, Beijing, China. ACL.

From *Frying* to *Speculating*: Google Ngram evidence to the meaning development of ‘炒’ in Mandarin Chinese

Jing Chen, Chu-Ren Huang
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong
jing95.chen@connect.polyu.hk, churen.huang@polyu.edu.hk

Abstract

This paper explores semantic change simply using the Ngram information, with the intuition that Ngrams as the direct neighbors offer collocation cues that further signal meaning changes. We specifically investigate the case of ‘炒’ in Mandarin Chinese on the basis of unigrams extracted from *Google Books Ngrams Corpus* and reconstruct the meaning development of ‘炒’ with specific stages. The results indicated that the major meaning changes that occurred to ‘炒’ is from ‘frying’ to ‘speculating’, which roughly started in the 1970s. The attested word types related to the latter sense denote economic events, such as stocks, foreign currencies, and speculators. It further reflects that social context plays an essential role in the process of semantic change.

1 Introduction

The availability of large historical digital corpora and the recent advance in natural language processing have greatly facilitated empirical studies on semantic change nowadays (Michel et al., 2011; Hamilton et al., 2016a,b; Tahmasebi et al., 2019). Beyond the world of historical linguistics, the computational community has shown a growing interest in exploiting statistical and computational models to automatically detect semantic change over the past two decades, from monitoring the fluctuation in the frequency of target words in historical texts (Michel et al., 2011; Hilpert and Gries, 2009; Kulkarni et al., 2014) to measuring their context differences using the state-of-art distributional models(Kim et al., 2014; Hamilton et al., 2016b; Giulianelli et al., 2020).

Building off of the distributional hypothesis in linguistics, ‘*You shall know a word by its company*’ (Firth, 1957), the distribution-based approach has taken up the predominant position in the recent lexical semantic change detection task (Schlechtweg et al., 2020; Tahmasebi et al., 2019). In general, distributional differences over a period of time could be quantified by constructing and evaluating historical word embeddings (Hamilton et al., 2016a,b). However, this novel trend is still in its youth. Most existing studies focus on evaluating how much the overall distribution of a word form deviated over two or three intervals but tell little about how the meaning of a word developed as a continuous process (Kutuzov and Pivovarova, 2021; Rodina and Kutuzov, 2020; Tahmasebi et al., 2019; Kutuzov et al., 2018). More importantly, the popular word-type models, also known as static word embeddings, may not be sensitive to less contrastive usage drifts (Tahmasebi et al., 2019; Schlechtweg et al., 2020).

In this paper, we present a simple and interpretable way to track the meaning development by collecting evidence directly from N-grams data extracted from *Google Books Ngram* dataset(the Chinese subpart) and manually checking the direct neighbours of target words over times. The intuition here is that N-grams provide collocation information and its changes may signal semantic changes. Besides, the collocations are also more linguistically informative in the sense of interpreting the specific stages of meaning development. Based on this intuition, we discuss the meaning development of ‘炒’ in Mandarin Chinese as a case study and find that it develops a sense of ‘speculating’ from the original ‘frying’ sense over the observed period, which coincides but supplements existing observations with empirical data (Diao, 1995; Shen, 2009).

The remainder of this paper is organized as follows. Section 2 summarizes related work and then situates our study. In Section 3, we introduce how the Ngrams data for this study have been collected. Section 4 presents the distribution information of ‘炒’ in the collected data, and Section 5 discusses the possible changing path of ‘炒’ and its possible reasons.

2 Related work

Semantic change generally refers to ‘a form historically acquires a new function to replace or augment its old ones’ (Sweetser, 1990). Studies anchored in the distributional hypothesis where ‘difference of meaning correlates with difference of distribution’ (Harris, 1954) assume that meaning change could be quantified by its neighboring information over time. For example, broadening and narrowing, regarded as two fundamental categories of meaning change, could be further interpreted as extending or contracting context varieties for target words (Campbell, 2013).

Following this working hypothesis, most recent studies use computational methods to monitor the change of context varieties to detect meaning change. Sagi et al. (2009, 2011) performed semantic density analysis by measuring the average cosine similarities of context vectors to identify the increase and decrease of context dispersions. The density information would be further interpreted as the general gain or loss of senses.

Tang et al. (2013, 2016) and Schlechtweg et al. (2017) exploited the *entropy* concept in Information theory to measure the gain or loss of information for target words. For example, Schlechtweg et al. (2017) specifically detected the metaphorical changes in German, an analogical mapping process from a more ‘concrete’ source domain onto a more ‘abstract’ target domain (Traugott and Dasher, 2001). They made use of entropy as an indicator to quantify the semantic generality of target words, which further calculate the meaning change.

With the very recent development in computational science, contextualized word embeddings have also been exploited in the detection task (Devlin et al., 2019; Hu et al., 2019; Giulianelli et al., 2020). For example, Giulianelli et al. (2020) first used K-Means clustering to group word token representations derived from BERT models into different word usage types and then applied the Entropy difference and the Jensen-Shannon divergence metrics to measure variations in the relative prominence of coexisting usage types.

The above proposals showed inspiring future directions for automatically detecting semantic change. However, training historical word embeddings, and especially clustering token embeddings have requirements on the computing capacity. The cumulative nature of meaning development and sophisticated relations among senses pose huge challenges for both statistical and neural language models. More crucially, distributional representations have a notorious bias on the frequency of target words. For example, a lower frequency of a novel sense may not be salient enough to be detected.

In this paper, we investigate the meaning development of ‘炒’ by checking its direct neighbors in the Google Ngram corpus. N-grams refer to an n-word sequence (Jurafsky and Martin, 2009), which provides neighboring information for the target words. To some extent, N-grams also reflect collocation preferences. It is clear that the change of collocation preferences would be a more precise and interpretable indicator for meaning development.

3 Method

Google Books Ngram Corpus is a collection of digitized books with over 500 billion words in 7 languages, which is publicly accessible [1]. The data in the corpus is stored in the n-grams format(n is with a maximum to 5) in order to protect intellectual property, containing ngrams with its frequency information in each specific year (Michel et al., 2011; Lin et al., 2012).

For our investigation, we used the Chinese (simplified) subcorpus(version 2), with texts spanning from 1990 to 2012. We first exploit python to crawl all 1-gram and 2-grams (together with their frequencies in each specific year) that occurred larger than two times in the corpus and then filter in words containing the ‘炒’ character in 1-gram and 2-grams for further analysis.

After manually checking all 1-gram data and 2-grams data containing ‘炒’, we found that unigram data actually provide sufficient cues for depicting the meaning development of the case ‘炒’. Highly frequent collocations with the sense of ‘speculating’, such as ‘炒股’(speculate in stocks) and ‘炒汇’(speculate in foreign currencies) and have been segmented as one unit, that is, 1-gram, over the

[1] Google Books Ngram dataset could be accessed via: https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

whole observed period. Likely, highly frequent expressions with the sense of 'frying', such as '炒米' (fried rice), '炒面'(fried noodles), are also segmented as unigrams.

According to Modern Chinese Dictionary(the 7th edition) (Department of Chinese Lexicography, 2019)，'炒' has four senses: 1) to fry; 2-3) used as '炒作', to speculate or hype; 4) get fired. As mentioned above, highly frequent collocations associated with senses 2-4 have been conventionalized as one unit, such as '炒作', '炒鱿鱼'. We thus assume that 1-gram for '炒' containing collocation and temporal information provides possibilities for meaning reconstruction.

4 The meaning distribution of '炒'

In the 1-gram data, there are 10 word types containing the character '炒'(frying, speculating): '炒米' (fried rice), '炒鱿鱼' (get fired), '炒家'(speculator), '炒汇'(to speculate in foreign currency), '炒面'(fried noodles), '炒货'(fried snacks), '炒股'(to speculate in stocks), '炒冷饭'(to heat leftover, to repeat without any new content), '炒作'(speculating, hype).

Among these words, '炒作' refers to 'speculate' and 'hype', which could be further regarded as less specific action compared with '炒股' etc. We first surveyed '炒' and '炒作' in the Google Ngram Viewer. As Figure 1 below shows, '炒' appeared at least before 1900, while '炒作' was first attested in 1979. A sharp increase of '炒作' occurred in the 1990s.

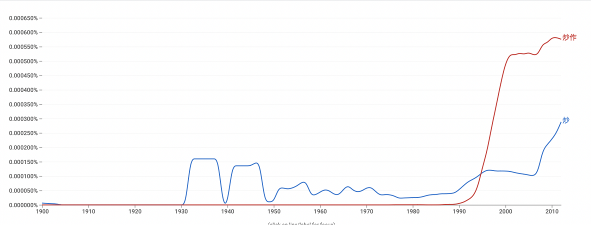


Figure 1: '炒' and '炒作' in *Google Ngram Viewer*

We plotted the remaining 8 word types using the raw data extracted from the corpus to discuss their distributions over time (see Figure 2 below). Among these words, '炒鱿鱼', '炒家', and '炒股' demonstrated a significant rise after the 1980s approximately. In contrast, '炒面', '炒米', '炒冷饭', and '炒货' stay relatively stable during the period from 1900 to 2012.

We also acquired their distributions in *Google Books Ngram Viewer*, which are plotted after normalization and smoothing. As illustrated in Figure 3, frequencies of '炒股', '炒家', '炒汇' have surged around the 1990s, while '炒货', '炒面', '炒冷饭' are much stable [2].

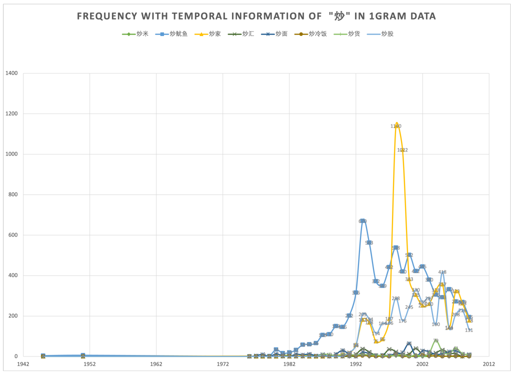


Figure 2: '炒' in 1gram: word type, raw frequency, and temporal information

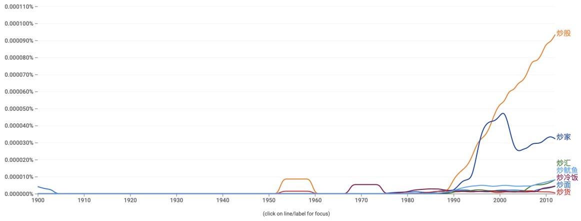


Figure 3: '炒面', '炒货', '炒冷饭', '炒股', '炒汇', '炒家', '炒鱿鱼' in *Google Books Ngram Viewer*

5 Reconstructing the meaning development of '炒'

The frequency and its temporal information regarding the distributions over time provide evidence to trace the meaning development. Data derived from Google Books Ngram indicates the general path of meaning development of '炒', from frying to speculating. As seen from the above figures, '炒' denoting cooking-related action appeared much earlier than the one representing abstract action (such as speculating in stocks and foreign currencies). The latter sense was first attested in the late 19th century. These two senses are closely related, such as having common characteristics of' fast stirring'.

One interesting case '炒作' have senses of 'to speculate' and 'to hype', frequently collocated as '炒作股票'(to speculate in stocks) and '炒作新闻'(to hype). We search these collocations in

[2] The google Ngram viewer uses normalized data for plotting. The low raw frequency of '炒米' made it invisible in the plot.

Google Ngram Viewer (see Figure 4). ‘炒作’ was first attested in the late 1970s, ‘炒作股票’ was first attested in the late 1980s, and ‘炒作新闻’ or ‘新闻炒作’ was in the early 1990s.

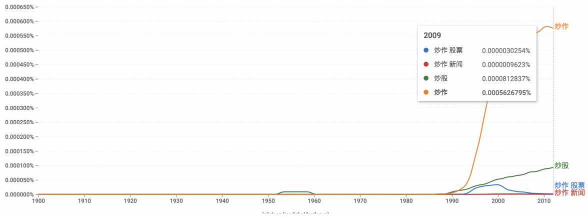


Figure 4: ‘炒作’, ‘炒股’, ‘炒作新闻’, ‘炒作股票’ in *Google Books Ngram Viewer*

Compound verbs such as ‘炒米’, ‘炒面’, ‘炒冷饭’, and ‘炒鱿鱼’ all have a basic sense of ‘frying’. ‘炒鱿鱼’ and ‘炒冷饭’ later acquired extra extended meanings, respectively. For ‘炒冷饭’, ‘to repeat without new content’ was metaphorically developed based on the specific event ‘to heat leftovers’. Similarly, ‘炒鱿鱼’, primarily denoting ‘get fired’ now, also got developed based on the original sense ‘fried squid’. In contrast, ‘炒股’, ‘炒汇’, and ‘炒家’ are closely related to ‘to speculate’, much deviated from the basic sense of ‘frying’.

According to the above analysis, we roughly reconstruct the meaning evolution of ‘炒’ (see Figure 5). The main changing path of ‘炒’ is from frying to speculating, and this change approximately started in the 1970s as witnessed that related compound words with a sense of speculating had a surge in terms of frequencies. The rising tendency even becomes more salient around the 1990s. The latter sense is closely related to economic events, such as stocks, foreign currencies, and speculators, in our data. The development of the ‘speculating’ sense roughly coincides with the timeline of the Reform and Opening-up, one of the most influential milestones in the recent history of modern China. It is generally assumed that this remarkable social change brought significant changes to the lexicon of Modern Chinese (Diao, 1995; Lin, 2021). In this case, we would assume that profound social change is one of the most fundamental driving forces behind the meaning development of ‘炒’.

6 Conclusion

In this paper, we reconstructed the meaning development for ‘炒’ in Mandarin Chinese, using

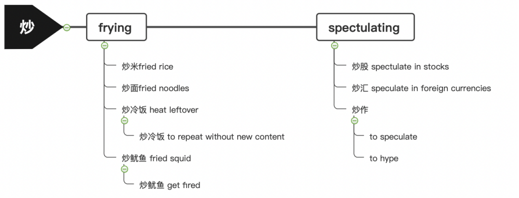


Figure 5: A possible changing path for the meaning of ‘炒’

Google Ngrams data, with the intuition that collocation information in Ngrams helps detect meaning change. Our results also provided more specific time information for the stages when ‘炒’ acquired the sense of ‘to speculate’. The meaning development of ‘炒’, from denoting the concrete cooking-related action to a relatively abstract economics-related action, further reflected that the social context plays an essential role in semantic change.

This paper showcased that Ngrams provide possibilities to depict the changing path of meaning directly. However, there are also some limitations in this study. For example, given the nature of Ngrams, a sliding window for a given sentence or sequence, there are ngrams that are less informative in terms of collocation. A question here is about how to evaluate ngrams in terms of its collocational weights. Another related question is about which types of target words would be suitable for such detection. These questions are served as research directions in the near future.

References

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

Chinese Academy of Social Science Department of Chinese Lexicography, Institute of Linguistics. 2019. *Contemporary Chinese Dictionary (Xiandai Hanyu Cidian)*, the 7th edition. Commercial Press, Peking.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Yanbin Diao. 1995. *The Development and Reform of Mainland Chinese in the New Era*. Hung Yeh Publishing, Taibei.

J.R. Firth. 1957. *A Synopsis of Linguistic Theory, 1930-1955*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of ACL*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of EMNLP*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Martin Hilpert and Stefan Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24:385–401.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the ACL International Workshop on Computational Approaches to Historical Language Change*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of COLING*.

Sheng Lin. 2021. *30 Words that have experienced "expressional" changes in recent years*, pages 369–380. De Gruyter Mouton, Berlin, Boston.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, and Peter Norvig. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: A Dataset of Historical Lexical Semantic Change in Russian. In *Proceedings of COLING*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL Workshop on GEMS: Geometrical Models of Natural Language Semantics*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. *Tracing semantic change with Latent Semantic Analysis*, pages 161–183.

Dominik Schlechtweg, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 354–367, Vancouver, Canada. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of SemEval*.

Mengying Shen. 2009. *New Words and New Expressions in Chinese New Era (1949-2009)*. Sichuan Lexicographical Press, Chengdu.

E. Sweetser. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of Computational Approaches to Lexical Semantic Change. *arXiv preprint arXiv:1811.06278*.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. Semantic Change Computation: A Successive Approach. In *Behavior and Social Computing*, pages 68–81, Cham. Springer International Publishing.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic Change Computation: A Successive Approach. *World Wide Web*, 19.

E.C. Traugott and R.B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

A Model-Theoretic Formalization of Natural Language Inference Using Neural Network and Tableau Method

Ayahito Saji[1], **Yoshihide Kato**[2], **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University
[2]Information and Communications, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
saji.ayahito.y7@s.mail.nagoya-u.ac.jp

Abstract

Saji et al. (2021) integrated a neural-based NLI model and a symbolic one into a framework to get the best of both worlds. This framework is based on a tableau method, which is a proof system for formal logic; however, it has a remaining issue that its theoretical limitations have not been clarified. To solve this issue, this paper formalizes the framework model-theoretically. On the basis of the formalization, we demonstrate that a certain kind of soundness holds for this framework, while the completeness does not.

1 Introduction

The natural language inference (NLI) (Dagan et al., 2013) is the task of identifying the inferential relation between a text pair: a *premise* and *hypothesis*. If a hypothesis can be inferred from a premise using logical and commonsense knowledge, it is judged as *entailment*. If a premise and hypothesis are incompatible, it is judged as *contradiction*, and if neither of these cases hold, it is judged as *neutral*. NLI systems are expected to be applied to a wide range of fields, e.g., question answering, information retrieval, and text summarization.

In recent years, neural-based approaches (Parikh et al., 2016; Chen et al., 2017; Devlin et al., 2019; Lan et al., 2019; Tai et al., 2020; Wang et al., 2021) have achieved high accuracy in experiments with many NLI datasets, e.g., the SNLI corpus (Bowman et al., 2015), MultiNLI (MNLI) corpus (Williams et al., 2018), Adversarial NLI (ANLI) dataset (Nie et al., 2020), and QNLI dataset (Wang et al., 2018). However, neural-based models have the limitation that they cannot explain the reasoning processes by which inferential relations are derived. Such models represented a black box, and it is difficult to analyze what kind of inference was performed.[1] Yanaka et al. (2020) proposed a method to evaluate whether neural models learn the systematicity of monotonicity inference in natural language, and they demonstrated that the generalization ability of current neural models is limited. Gururangan et al. (2018) and Tsuchiya (2018) showed that NLI datasets such as the SNLI corpus and MNLI corpus have a hidden bias in that only a hypothesis is sufficient to determine inferential relations, and there is a risk that neural models are simply identifying inferential relations based on such biases.

From a different perspective, symbolic approaches to the NLI task have been proposed (Bar-Haim et al., 2007; MacCartney and Manning, 2007, 2008, 2009; Mineshima et al., 2015; Abzianidze, 2015, 2017; Hu et al., 2020). These approaches have the advantage that the reasoning process deriving the relation is understandable to humans, unlike neural-based approaches. In addition, symbolic approaches are generally founded on formal logic or linguistic analyses, which allows us to understand the reasoning processes. However, these approaches require inference rules that are created by humans; thus, it is difficult to handle synonyms, hypernyms, and hyponyms exhaustively. In addition, commonsense knowledge, e.g., “when it rains, the ground gets wet” must be expressed as inference rules; however, it is unlikely that such rules can be created exhaustively. Currently, these approaches have only been tested for the controlled NLI datasets such as the FraCaS test suite (Cooper et al., 1996) and SICK dataset (Marelli et al., 2014).

Saji et al. (2021) integrated these two approaches into a unified framework that can make the inference process explicit for certain linguistic phenomena while maintaining the applicability of neural-based approaches to relatively freely created (non-controlled) datasets. This method is based on the tableau method, which is one of the proof methods of formal logic; however, to the best of our

[1] One of the exceptions is the NLI system, which generates explanations by Kumar and Talukdar (2020); however, since they are generated by the neural model, the process of generating explanations is a black box.

knowledge, its theoretical limitations have not been clarified.

Thus, in this paper, we formalize the method proposed by Saji et al. model-theoretically to clarify the theoretical limitations of the method. We demonstrate that a certain kind of soundness holds for this method, while completeness does not.

2 Natural Language Inference using Neural Network and Tableau Method

This section provides an overview of the method proposed by Saji et al. For a premise $p \in L$ and hypothesis $h \in L$,[2] the method takes the following steps:

1. Construct the tableaux to prove entailment and contradiction relations.
2. Judge the closedness of the tableaux using a neural-based NLI system.
3. Determine the relation between the premise and hypothesis based on the constructed tableaux.

Below, we explain each step.

2.1 Tableaux Construction

The tableau method constructs a tree structure, referred to as a *tableau*, for a given set, each element of which is a pair of a sentence s and truth value v. Each node in a tableau is labeled with a tuple $e = (s, v, a, o)$ called an *entry*. Here, the entry e represents the constraint that s must take v as its truth value. a is the flag indicating whether the tableau rule (described below) has been applied to e, and o denotes the entry from which e was derived. The truth value v is either $\mathtt{T}$ or $\mathtt{F}$, representing true or false, respectively. The flag a is either 0 or 1, where 0 means that no tableau rule has been applied and 1 means that some rule has been applied. The *initial tableau* consists of entries for a given pair of natural language sentence and truth value, with $a = 0$. Each entry in the initial tableau is assumed to be derived from itself. That is, for each entry e in the initial tableau, e is in the form of $(s, v, 0, e)$. The tableau is created by repeatedly applying *tableau rules* to the entries. By applying a tableau rule to an entry, the constraint expressed by the entry is decomposed into several constraints. The decomposed constraints are added to the tableau as new entries. This decomposition makes the reasoning process explicit. In the following, when a tableau t' is derived by applying a tableau rule $r \in R$ to a tableau t, we write $t \overset{R}{\triangleright} t'$, and we refer to a tableau t which has no tableau t' such that $t \overset{R}{\triangleright} t'$ as a *complete tableau*.

Branches in a tableau indicate that there are multiple cases for possible valuation. A complete tableau to prove entailment relation whose initial tableau consists of $\{e_1 = (p, \mathtt{T}, 0, e_1), e_2 = (h, \mathtt{F}, 0, e_2)\}$ is referred to as an *entailment tableau* and that for contradiction whose initial tableau consists of $\{e_1 = (p, \mathtt{T}, 0, e_1), e_2 = (h, \mathtt{T}, 0, e_2)\}$ is referred to as a *contradiction tableau*.

A tableau rule takes the following form:[3]

$$r = (c_{1,1} \wedge \cdots \wedge c_{1,n_1}) \vee \cdots \vee (c_{m,1} \wedge \cdots \wedge c_{m,n_m}),$$

where $c_{1,1}, \ldots, c_{1,n_1}, \ldots, c_{m,1}, \ldots, c_{m,n_m}$ are functions that take $(s, v, 0, o)$ as input and return $(s', v', 0, o)$. Here, if $c_{i,j_i}(e)$ is defined for all c_{i,j_i} a tableau rule r is *applicable* to e. If a tableau has an entry e to which r is applicable, new m branches $\langle c_{1,1}(e), \cdots, c_{1,n_1}(e)\rangle, \cdots, \langle c_{m,1}(e), \cdots, c_{m,n_m}(e)\rangle$ are added as their children to all leaves dominated by the entry e. A tableau rule converts the constraint expressed in the source entry into the equivalent one.[4] There is no need to apply an operation to the entry e to which the tableau rule has been applied, because the constraint is already expressed by the entries derived from the entry e. The flag a in an entry (s, v, a, o) controls the application of such operations and is changed to 1 when a tableau rule is applied. An entry $(s, v, 1, o)$ is neither applied any tableau rules nor used in judging the closedness of the tableau as described below. Figure 1 (left) shows an entailment tableau for the following example:

Premise Either Smith or Anderson signed the contract.

Hypotheses If Smith did not sign the contract Anderson made an agreement.

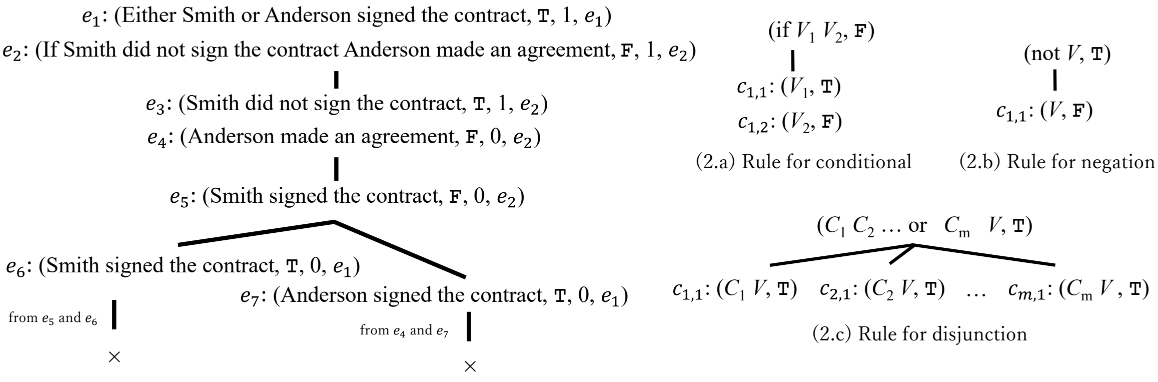
Label Entailment

The initial tableau of this example consists of e_1 = (Either Smith or Anderson signed the contract, $\mathtt{T}$, 0, e_1) and e_2 = (If Smith did not sign

[2] L is the set of natural language sentences.

[3] For a specific implementation of the tableau rules, see (Saji et al., 2021).

[4] Although tableau rules that derive weak constraints are also possible, such rules are not considered in the method proposed by Saji et al.


e_1: (Either Smith or Anderson signed the contract, T, 1, e_1)
e_2: (If Smith did not sign the contract Anderson made an agreement, F, 1, e_2)
e_3: (Smith did not sign the contract, T, 1, e_2)
e_4: (Anderson made an agreement, F, 0, e_2)
e_5: (Smith signed the contract, F, 0, e_2)
e_6: (Smith signed the contract, T, 0, e_1)
e_7: (Anderson signed the contract, T, 0, e_1)
from e_5 and e_6
×
from e_4 and e_7
×
(if V_1 V_2, F)
$c_{1,1}$: (V_1, T)
$c_{1,2}$: (V_2, F)
(not V, T)
$c_{1,1}$: (V, F)
(C_1 C_2 … or C_m V, T)
$c_{1,1}$: (C_1 V, T)
$c_{2,1}$: (C_2 V, T)
…
$c_{m,1}$: (C_m V, T)

(2.a) Rule for conditional

(2.b) Rule for negation

(2.c) Rule for disjunction

(1) Example of entailment tableau

(2) Examples of tableau rules

Figure 1: Example of entailment tableau in Saji et al.'s method

the contract Anderson made an agreement, F, 0, e_2). First, applying the rule (2.a) of Figure 1 to e_2 adds two new entries at the end of the path (the tableau leaf): $e_3 =$ (Smith did not sign the contract, T, 0, e_2) and $e_4 =$ (Anderson made an agreement, F, 0, e_2), and the flag of e_2 is changed to 1. Second, applying the rule (2.b) to e_3 adds a new entry: $e_5 =$ (Smith signed the contract, F, 0, e_2) and the flag of e_3 is changed. Finally, applying the rule (2.c) to e_1 adds two new entries: $e_6 =$ (Smith signed the contract, T, 0, e_1) and $e_7 =$ (Anderson signed the contract, T, 0, e_1), and the flag of e_1 is changed.

2.2 The Closedness of Tableaux

Saji et al. defined a branch b is *closed*, if and only if two entries $e_1 = (s_1, v_1, 0, o_1)$ and $e_2 = (s_2, v_2, 0, o_2)$ $(o_1 \neq o_2)$ on b satisfy one of the following three conditions, which we refer to as *closedness conditions*:

1. $v_1 = \mathtt{T} \wedge v_2 = \mathtt{F} \wedge s_1 = s_2$
2. $v_1 = \mathtt{T} \wedge v_2 = \mathtt{T} \wedge NLI(s_1, s_2) = \mathtt{C}$
3. $v_1 = \mathtt{T} \wedge v_2 = \mathtt{F} \wedge NLI(s_1, s_2) = \mathtt{E}$

Here, $NLI(s_1, s_2)$ is any NLI system that takes premise $s_1 \in L$ and hypotheses $s_2 \in L$ as inputs and returns one of the following classes: entailment (E), neutral (N), or contradiction (C). The first condition is similar to that of conventional tableau method, and the other two conditions are based on the NLI system. If all branches in a tableau are closed, the tableau is *closed*. The condition $o_1 \neq o_2$ excludes entries derived from only a premise or a hypothesis from the judging of the closedness; however, this is not a problem if a premise or hypothesis is neither tautology nor contradictory sentence.

Entailment tableau	Contradiction tableau	Output
Closed	Not closed	Entailment
Not closed	Closed	Contradiction
Not closed	Not closed	Neutral
Closed	Closed	Error

Table 1: Correspondence between closedness of tableaux and the inferential relation

As an example, let us consider the tableau shown in Figure 1 (left) and assume that $NLI\big(sen(e_7), sen(e_4)\big) = \mathtt{E}$.[5] The tableau has two branches. The left branch $\langle e_1, e_2, e_3, e_4, e_5, e_6\rangle$ is closed because e_5 and e_6 satisfy the first closedness condition. The right branch $\langle e_1, e_2, e_3, e_4, e_5, e_7\rangle$ is also closed because e_4 and e_7 satisfy the third closedness condition.

2.3 Determining the Inferential Relation

The inferential relation between a premise and a hypothesis is predicted based on the closedness of entailment and contradiction tableaux. This is identical to that of Abzianidze (2015) and summarized in Table 1.[6]

3 Model-Theoretic Formalization

3.1 Model

To discuss the method proposed by Saji et al. formally, we first define a model-theoretic interpretation of sentences. We then characterize tableau rules and NLI systems based on such interpretation, and define some properties of their proof system.

[5] Here, $sen\big((s, v, a, o)\big) = s$.

[6] The "error" class is for an uninterpretable situation where both entailment and contradiction relations hold.

A model is defined as follows:

Definition 1 (Model)**.** A model is a function from L to $\{\mathtt{T},\mathtt{F}\}$. Let $\mathcal{M}$ be the set of all models. For a set of models $M \subseteq \mathcal{M}$, we define $M(s) = \{m \in M \mid m(s) = \mathtt{T}\}$.

Intuitively, a set of models $M(s)$ can be considered a set of situations where s is true. In the following, we define what NLI systems and tableau rules are consistent with a given set of models M,[7] and we clarify the theoretical limitations of the tableau methods of Saji et al. under the condition where the NLI system and tableau rules are consistent with the model set.

Definition 2. An NLI system is said to be *consistent* with a model set M if and only if all of the following conditions hold.

- $M(p) \subseteq M(h) \Leftrightarrow NLI(p,h) = \mathtt{E}$
- $\neg(M(p) \subseteq M(h)) \wedge \neg(M(p) \cap M(h) = \emptyset) \Leftrightarrow NLI(p,h) = \mathtt{N}$
- $M(p) \cap M(h) = \emptyset \Leftrightarrow NLI(p,h) = \mathtt{C}$

In this definition, the entailment relation corresponds to the inclusion relation of the model set, and the contradiction relation as the relation that the model sets are mutually disjoint.

In the following, we characterize the tableau methods based on the model-theoretic interpretation. In preparation, we define the model set for an entry, a branch of the tableau and a tableau by extending $M(s)$.

Definition 3 (Model set for an entry)**.** For a tableau entry $e = (s, v, a, o)$, we define $M(e)$ as follows:

$$M(e) = \begin{cases} M(s) & (v = \mathtt{T}) \\ M - M(s) & (v = \mathtt{F}) \end{cases}$$

Definition 4 (Model set for a branch)**.** For a branch b of a tableau, we define $M(b)$ as follows:

$$M(b) = \cap_{e \in b} M(e).$$

Definition 5 (Model set for a tableau)**.** Let B be a set of all branches in a tableau t. We define $M(t)$ as follows:

$$M(t) = \cup_{b \in B} M(b).$$

Definition 6. Let $r = (c_{1,1} \wedge \cdots \wedge c_{1,n_1}) \vee \cdots \vee (c_{m,1} \wedge \cdots \wedge c_{1,n_m})$, E be the set of all entries, and $E_r = \{e \in E \mid r \text{ is applicable to } e\}$. A tableau rule r is said to be *consistent* with a model set M if and only if the following condition is satisfied for all $e \in E_r$:

$$M(e) = \Big(M\big(c_{1,1}(e)\big) \cap \cdots \cap M\big(c_{1,n_1}(e)\big)\Big) \cup \cdots \cup \Big(M\big(c_{m,1}(e)\big) \cap \cdots \cap M\big(c_{1,n_m}(e)\big)\Big).$$

In this equation, the left-hand side corresponds to the constraint of the entry e. The right-hand side represents the constraint derived by applying the tableau rule r to the entry e.

In addition, a tableau rule set R is said to be consistent with M if and only if all $r \in R$ are consistent with M.

3.2 Soundness and Completeness

The main components of the method proposed by Saji et al. are a tableau rule set R and NLI system NLI. We call a pair (R, NLI) *a proof system* and define the soundness and completeness of the proof system based on the model-theoretic interpretation. Furthermore, we prove any proof systems are sound and give a counterexample for the completeness. The soundness defined in this paper is the property that if the entailment (contradiction) tableau is closed, then the model sets for the premises and hypotheses are in an inclusion (mutually disjoint) relation. The completeness is the converse of the soundness. We define the soundness of a proof system as follows:

Definition 7 (Soundness)**.** Let M be a model set, R be a set of tableau rules consistent with M, and NLI be an NLI system consistent with M. We say that the proof system (R, NLI) is *sound* with respect to M if and only if the following condition holds for all premise p and hypothesis h:

- If the entailment tableau constructed by R is closed by NLI, then $M(p) \subseteq M(h)$.
- If the contradiction tableau constructed by R is closed by NLI, then $M(p) \cap M(h) = \emptyset$.

On the other hand, the completeness is defined as follows:

Definition 8 (Completeness)**.** Let M be a model set, R be a set of tableau rules consistent with M, and NLI be an NLI system consistent with M.

[7] The reason why we consider a subset M of $\mathcal{M}$ is that $\mathcal{M}$ contains logically unnatural models such as models that return $\mathtt{T}$ for every natural language sentence. We assume that the set of logically natural models is given as M. No conditions are required for M; thus, the following discussion is valid no matter what kind of M is.

We say that the proof system (R, NLI) is *complete* with respect to M if and only if the following condition holds for all premise p and hypothesis h:

- If $M(p) \subseteq M(h)$, then the entailment tableau constructed by R is closed by NLI.
- If $M(p) \cap M(h) = \emptyset$, then the contradiction tableau constructed by R is closed by NLI.

3.2.1 Proof of Soundness

In this section, we prove the following theorem:

Theorem 1 (Soundness Theorem)**.** Let M be a model set. Any proof systems (R, NLI) consistent with M are sound with respect to M.

First, we introduce two lemmas to prove the theorem.

Lemma 2. Let M be a model set and R be a tableau rule set consistent with M. The following equation holds for any tableaux t and t' such that $t \overset{R}{\rhd} t'$:

$$M(t) = M(t').$$

From Definition 6, Lemma 2 is trivial.

Lemma 3. Let t and t' be tableaux such that $t \overset{R^*}{\rhd} t'$. If R is consistent with M, $M(t) = M(t')$.

Lemma 3 is trivial from Lemma 2.

The proof of Theorem 1 is as follows:

Proof. Assume that the entailment tableau t_{ent} constructed by R is closed by NLI. This implies that, two entries $(s_1, v_1, 0, o_1)$ and $(s_2, v_2, 0, o_2)$ $(o_1 \neq o_2)$ satisfy one of the closedness conditions in all branches b of t_{ent}.

- If the first closedness condition is satisfied, $v_1 = \mathtt{T} \wedge v_2 = \mathtt{F} \wedge s_1 = s_2$. Thus, from Definition 3, $M(e_1) \cap M(e_2) = M\big((s_1, \mathtt{T}, 0, o_1)\big) \cap M\big((s_1, \mathtt{F}, 0, o_2)\big) = M(s_1) \cap \big(M - M(s_1)\big) = \emptyset$.
- If the second closedness condition is satisfied, we obtain $v_1 = \mathtt{T} \wedge v_2 = \mathtt{T} \wedge NLI(s_1, s_2) = \mathtt{C}$. NLI is consistent with M; thus, from Definition 2, $M(s_1) \cap M(s_2) = \emptyset$. Therefore, from Definition 3, $M(s_1) \cap M(s_2) = M\big((s_1, \mathtt{T}, 0, o_1)\big) \cap M\big((s_2, \mathtt{T}, 0, o_2)\big) = M(e_1) \cap M(e_2) = \emptyset$.
- If the third closedness condition is satisfied, we obtain $v_1 = \mathtt{T} \wedge v_2 = \mathtt{F} \wedge NLI(s_1, s_2) = \mathtt{E}$. NLI is consistent with M; therefore, from Definition 2, $M(s_1) \subseteq M(s_2)$. Thus, from Definition 3, $M(e_1) = M(s_1)$, $M(e_2) = M - M(s_2)$. Since $M(s_1) \subseteq M(s_2)$, $M(e_1) \cap M(e_2) = \emptyset$.

From the above, $M(e_1) \cap M(e_2) = \emptyset$. Thus, from Definition 4, $M(b) = \emptyset$. Here, $M(b) = \emptyset$ for all branches $b \in t_{\text{ent}}$; thus, from Definition 5, $M(t_{\text{ent}}) = \emptyset$. Let t_{init} be the initial tableau. By $t_{\text{init}} \overset{R^*}{\rhd} t_{\text{ent}}$ and Lemma 3, $M(t_{\text{init}}) = \emptyset$. Since t_{init} consists of $e_p = (p, \mathtt{T}, 0, e_p)$ and $e_h = (h, \mathtt{F}, 0, e_h)$, from Definition 4 and 5, $M\big((p, \mathtt{T}, 0, e_p)\big) \cap M\big((h, \mathtt{F}, 0, e_h)\big) = \emptyset$. Thus, $M\big((p, \mathtt{T}, 0, e_p)\big) \subseteq M\big((h, \mathtt{T}, 0, e_h)\big)$. From Definition 3, $M(p) \subseteq M(h)$. The above satisfies the first condition of Definition 7. The second condition (for the contradiction tableau) can be proved in a similar way. □

3.2.2 Counterexample of Completeness

In this section, we present a counterexample such that a proof system is not complete. Let us consider the following example:

Premise Smith and Anderson did not go out.

Hypothesis Smith and Anderson were home.

Label Entailment

The entailment tableau shown in Figure 2 is for the premise and hypothesis. The truth values of entries e_4, e_5, e_6 and e_7 to which the tableau rule has not yet been applied, are all $\mathtt{F}$. In the closedness conditions, the truth value of one of the entries must be $\mathtt{T}$; thus, this entailment tableau cannot be closed with any NLI. As an example of the model set M, let us consider the models shown in Table 2. In this model set M, $M(e_1) = M(e_3) = M(e_4) \cap M(e_5) = \{m_1\}$ and $M(e_2) = M(e_6) \cup M(e_7) = \{m_2, m_3, m_4\}$; thus, these tableau rules are consistent with M. Note that $M(p) \subseteq M(h)$ since $M(p) = M(h) = \{m_1\}$. This means that the first condition of Definition 8 does not hold. Thus, the completeness does not hold.

3.3 Example of Complete Proof System

Generally, the completeness does not hold; however, we can prove that a certain proof system is complete. In this section, we present an example of such proof system.

Let M be a model set, NLI be an NLI system that is consistent with M and R be a tableau rule

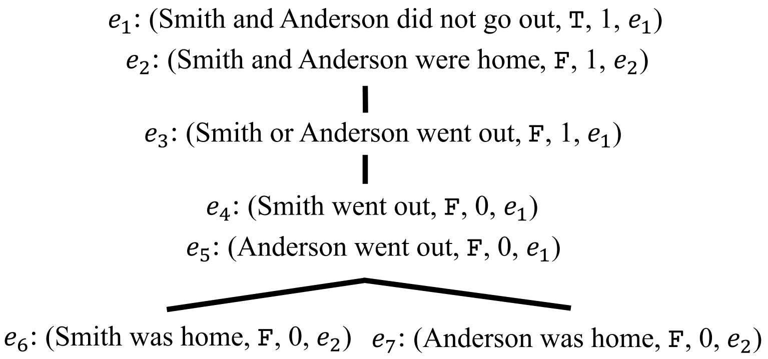

e_1: (Smith and Anderson did not go out, T, 1, e_1)
e_2: (Smith and Anderson were home, F, 1, e_2)
e_3: (Smith or Anderson went out, F, 1, e_1)
e_4: (Smith went out, F, 0, e_1)
e_5: (Anderson went out, F, 0, e_1)
e_6: (Smith was home, F, 0, e_2)
e_7: (Anderson was home, F, 0, e_2)

Figure 2: Counterexample of completeness

		m_1	m_2	m_3	m_4	M
Smith and Anderson did not go out	$\neg(\neg S \vee \neg A)$	T	F	F	F	$\{m_1\}$
Smith and Anderson were home	$S \wedge A$	T	F	F	F	$\{m_1\}$
Smith or Anderson went out	$\neg S \vee \neg A$	F	T	T	T	$\{m_2, m_3, m_4\}$
Smith went out	$\neg S$	F	F	T	T	$\{m_3, m_4\}$
Anderson went out	$\neg A$	F	T	F	T	$\{m_2, m_4\}$
Smith was home	S	T	T	F	F	$\{m_1, m_2\}$
Anderson was home	A	T	F	T	F	$\{m_1, m_3\}$

Table 2: Models for counterexample of completeness

set consistent with M such that for all $r \in R$, the following conditions hold:[8]

1. r is in the form of $c_{1,1}$.
2. If $c_{1,1}\big((s, v, 0, o)\big)$ is defined, the returned value is in the form of $(s', v, 0, o)$.

In this proof system, entailment tableaux constructed by R are always closed by NLI if $M(p) \subseteq M(h)$. The proof of its completeness is as follows:

Proof. Assume that $M(p) \subseteq M(h)$. Here, let t_{init} be the initial tableau constructed from $e_p = (p, \mathtt{T}, 0, e_p)$ and $e_h = (h, \mathtt{F}, 0, e_h)$. For all tableaux t such that $t_{\text{init}} \overset{R^*}{\rhd} t$, the following statements hold:

- t has only one branch.
- t has only two entries whose flag is 0: the one is in the form of $(p', \mathtt{T}, 0, e_p)$ and the other is in the form of $(h', \mathtt{F}, 0, e_h)$.
- $M(p) = M(p')$
- $M(h) = M(h')$

The first statement holds from the first condition about R. The second statement holds from the conditions about R. From the conditions about R, Definition 3 and 6, the last two statements hold. The entailment tableau t_{ent} derived from t_{init} also satisfies the above statements, since $t_{\text{init}} \overset{R^*}{\rhd} t_{\text{ent}}$. Because $M(p') = M(p) \subseteq M(h) = M(h')$ and NLI is consistent with M, $NLI(p', h') = \mathtt{E}$. Thus, the branch of t_{ent} satisfies the second closedness condition, that is, t_{ent} is closed by NLI. The second condition (for the contradiction tableau) of the completeness also can be proved in the similar way. Thus, this proof system is complete with respect to M. □

4 Conclusion

In this paper, we have formalized the method proposed by Saji et al. based on a model-theoretic interpretation and have clarified the theoretical limitations of this method. We have proved the soundness theorem and provided an example of complete proof system. In future work, we will explore what kind of proof systems are complete.

Acknowledgments

This research was partially supported by the Grant-in-Aid for Scientific Research (C) (No. 22K12148) of JSPS.

[8] We can consider that the tableau rules simply paraphrase sentences.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.

Lasha Abzianidze. 2017. LangPro: Natural language theorem prover. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 115–120, Copenhagen, Denmark. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Iddo Greental, and Eyal Shnarch. 2007. Semantic inference at the lexical-syntactic level. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 871–876.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. 2020. MonaLog: a lightweight system for natural language inference based on monotonicity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 334–344, New York, New York. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 4885–4901, Online. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Ayahito Saji, Daiki Takao, Yoshihide Kato, and Shigeki Matsubara. 2021. Natural language inference using neural network and tableau method. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 406–414, Shanghai, China. Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *ArXiv*, abs/2104.14690.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

Word Sense Disambiguation of Corpus of Historical Japanese Using Japanese BERT Trained with Contemporary Texts

Kanako Komiya
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho Koganei
Tokyo Japan 184-8588
kkomiya@go.tuat.ac.jp

Nagi Oki
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho Koganei
Tokyo Japan 184-8588
s182739s@st.go.tuat.ac.jp

Masayuki Asahara
National Institute for Japanese Language and Linguistics
10-2 Midoricho Tachikawa
Tokyo Japan 190-0014
masayu-a@ninjal.ac.jp

Abstract

Diachronic adaptation for word sense disambiguation (WSD) in Historical Japanese is known as a difficult problem because the most frequent sense (MFS) baseline is hard to beat. However, this paper reports that the model using BERT trained with contemporary texts significantly outperforms the MFS baseline. We also showed the effectiveness of multitask learning of WSD and document classification. We conducted the experiments using two sets of a sense-tagged corpus, Corpus of Historical Japanese sense-tagged with Word List by Semantic Principles: the datasets developed until 2019 and 2022. Finally, we discuss the reason why the diachronic adaptation for WSD of historical Japanese using BERT trained with contemporary Japanese is effective.

Introduction

Word sense disambiguation (WSD) involves the task of identifying the senses of words in documents. There have been a number of studies on WSD of Contemporary Japanese using sense-tagged corpora. However, due to the limitation of the sense-tagged corpora, it was difficult to achieve high performance for WSD of historical Japanese. To alleviate this problem, diachronic adaptation using contemporary Japanese has been tried for WSD of historical Japanese. However, the prior work shows that the most frequent sense (MFS) baseline is hard to beat for conventional methods that used examples from contemporary corpora and/or word embeddings trained with contemporary texts in addition to historical texts (Tanabe, 2020).

Meanwhile, Bidirectional Encoder Representations from Transformers (BERT) Devlin et al., (2019) substantially improved the state of the art of tasks of natural language processing, including WSD (Blevins and Zettlemoyer, 2020; Loureiro and Camacho-Collados, 2020). First, in this paper, we will show that WSD of historical Japanese using BERT significantly outperforms the MFS baseline and conventional methods. As BERT is trained mostly with contemporary Japanese texts (Japanese Wikipedia), WSD of historical Japanese using BERT is considered as a form of diachronic adaptation. Next, we tried multitask learning of

WSD and document classification using a sense-tagged corpus, Corpus of Historical Japanese (CHJ) [1](see Section 3).

We conducted the experiments using two sets of CHJ, the datasets developed until 2019 and 2022 (see Sections 4 and 5). Finally, we discuss why BERT is effective for WSD of historical Japanese in Sections 6 and 7.

The contributions of this paper are listed as follows:

(1) We show that the diachronic adaptation using BERT trained with contemporary Japanese substantially outperformed the MFS baseline and conventional methods for WSD of historical Japanese,

(2) We show that multitask learning of WSD and document classification of the text where the target word of WSD was taken from is effective when large amounts of training data was used, and

(3) We discuss what kind of information contributed to the diachronic adaptation for WSD of historical Japanese.

Related Work

WSD has two categories: lexical sample task and all-words WSD. Lexical sample task targets frequent words in a dataset (Iacobacci et al., 2016; Okumura et al., 2010; Komiya and Okumura, 2011) and all-word WSD disambiguates all words in a corpus (Raganato et al., 2017a; Shinnou et al., 2017; Iacobacci et al., 2016; Suzuki et al., 2018; Raganato et al., 2017b; Blevins and Zettlemoyer, 2020; Loureiro and Camacho-Collados, 2020). There have been a number of studies on WSD of contemporary Japanese of both categories. This paper focuses on the lexical sample task.

In addition, there have been some studies on historical Japanese texts. Hoshino et al. (2014) proposed translating historical Japanese to contemporary Japanese using a statistical machine translation system trained with a corpus obtained by their method using sentence alignment. Takaku et al. (2020) employed neural machine translation for translation from historical Japanese to contemporary Japanese. They used word embeddings diachronically fine-tuned with historical corpora, including word embeddings gradually fine-tuned in the order of time, which is proposed in Kim et al. (2014), for the input to their system and showed the fine-tuned word embeddings improved the translation performances.

Tanabe (2020) used the diachronically fine-tuned word embeddings for the WSD task including those trained following methods used by Takaku et al. (2020). According to (Daumé III et al., 2010; Daumé III, 2007), there are three types of approaches for domain adaptation depending on the information to be learned, namely, supervised, semi-supervised and unsupervised approaches. Tanabe (2020) used not only sense-tagged corpora but also unlabeled texts for diachronic adaptation, in three scenarios including all three types of domain adaptation approaches.

Related to WSD of historical Japanese, Tanabe et al. (2018) proposed a system to classify the word senses of words in a Japanese historical corpus to determine the word senses that are not listed in a dictionary of contemporary Japanese. However, they did not perform the WSD of historical Japanese itself.

This research is also related to the methods to capture the change of meanings. Kulkarni et al. (2015), Hamilton et al. (2016b), and Hamilton et al. (2016a) have shown the effectiveness of distributional semantics for this task. Kobayashi et al. (2021) used the BERT model and Aida et al. (2021) used PMI and SVD joint learning to capture the change of meaning of modern and contemporary Japanese.

Diachronic Adaptation Using BERT Trained with Contemporary Japanese Texts

As mentioned above, WSD of historical Japanese using BERT trained with contemporary Japanese texts is considered as a form of diachronic adaptation. We show that the method using the BERT model outperforms the MFS baseline and the conventional methods using word embeddings proposed by Tanabe (2020).

In addition, we attempted multitask learning of WSD and document classification. For the multitask learning of WSD and document classification, we simultaneously predicted not only word senses but also the literature the input

[1] https://clrd.ninjal.ac.jp/chj/overview-en.html

sentence was taken from. The motivation behind this method is to capture the diversity of the periods when each literature was written. We process all the historical Japanese texts at one time but the word sense in very old literature, say the work written in the 900s, should be different from that in relatively new literature like the work written in the 1600s. We also anticipated that the frequent senses vary depending on the literature work the input sentence was taken from.

Data

We used Corpus of Historical Japanese sense-tagged with Word List by Semantic Principles (CHJ-WLSP) (Asahara et al., 2022). Word List by Semantic Principles (WLSP) (National Institute for Japanese Language and Linguistics, 1964) is a Japanese thesaurus. In the WLSP, the article numbers or concept numbers indicate shared synonyms. In the WLSP thesaurus, words are classified and organized by their meanings. We can use the article numbers in WLSP with words as word senses. For example, the word “犬” (inu, meaning spy or dog) has two records in the WLSP, and therefore has two article numbers, 1.2410 and 1.5501, indicating that the word is polysemous. We can also use the historical version of WLSP (Miyajima et al., 2014).

We conducted the experiments using two sets of CHJ-WLSP, the datasets developed until 2019 and 2022. First, for comparison, we used the same data as (Tanabe, 2020), CHJ-WLSP developed until 2019. We refer to this data as CHJ-WLSP 2019. The literature was 5 works, that is, Taketori-monogatari (The Tale of the Bamboo Cutter), Tosa Nikki (Tosa Diary), Hōjōki (Square-jō record), Tsurezuregusa (Essays in Idleness), and Toraakira-bon Kyogen. Following (Tanabe, 2020), we used 58 words for the target words of WSD. They were selected because they appeared 50 times or more in CHJ and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014).

For the second experiments, we used data from 10 works, CHJ-WLSP developed until 2022 for this experiment. We refer to this data as CHJ-WLSP 2022. The additional literature was Konjaku Monogatarishū (Anthology of Tales from the Past), Jikkinsyo, Uji Shūi Monogatari, Taiyo magazine, and Kokutei textbook. We used 33 words, which are the words appeared more than 1,000 times in CHJ, for the target words of WSD. Table 1 displays the book title, number of word tokens, period, style of sub-corpora from CHJ we used for the experiments. Table 2 shown as an appendix lists the words, pronunciation, translation, and developed year of the data (2019 or 2022). Both in the table means the words are used for both experiments of CHJ-WLSP 2019 and CHJ-WLSP 2022. The translations shown in the table are just examples because the words are polysemous.

Book Title	Word Tokens	Period	Style
Taketori Monogatari (The Tale of the Bamboo Cutter)	12,757	Around 900	Fictional prose narrative
Tosa Nikki (Tosa Diary)	8,208	934	Poetic diary
Hōjōki (Square-jō record)	5,402	1212	Essay
Tsurezuregusa (Essays in Idleness)	40,834	1336	Essay
Toraakira-bon Kyogen	5,448	1642	Kyogen (Traditional theater)

Konjaku Monogatarishū (Anthology of Tales from the Past)	175,598	1100
Uji Shūi Monogatari	120,705	1220
Jikkinsyo	90,177	1252
Taiyo Magazine	46,394	1895-1
Kokutei Textbook	154,955	1910

Table 1 Book Titles, Number of Word Tokens, Period, Style of Books in CHJ of 5 works (CHJ-WLSP 2019) and 10 works (CHJ-WLSP 2022)

Experiments

We used bert-base-japanese-whole-word-masking, Japanese BERT mostly trained with contemporary Japanese texts (Japanese Wikipedia)[2], from transformers library.

For the simple BERT model, we used fine-tuning of BERT model added one layer. The input of the final layer is the output vector of the BERT model for the target word of WSD. We used softmax and cross entropy loss for the last layer. Stochastic Gradient Descent was used for the optimization function.

For the multitask learning, two final layers were added in parallel to the BERT model. One is for WSD, and another is for document classification. The input of the final layers is the same for the two final layers: the output vector of the BERT model for the target word of WSD.

Because we adopted the lexical sample task, a model is trained for each target word type of WSD. In other words, we prepared 58 models or 33 models for each method. The input of the system is sentence-based, and an example includes a target word token. Because of the length limitation of the BERT input, a sentence beyond 512 tokens is shorted to 512 tokens. In addition, if the target word of WSD was appeared in the omitted part of the sentence, we did not use the example for the experiments.

Mostly, we used Japanese period marks for dividing the sentences, but only for Toraakira-bon Kyoken, we used blank marks and boundary marks in CHJ to divide the sentences because it included no period marks. Tables 3 summarizes the maximum, minimum, and average number of data points for each target word of WSD.

	CHJ-WLSP 2019	CHJ-WLSP 2022
Max	419	7,072
Min	50	1,009
Avg.	167.83	2,255.15

Tables 3 The maximum, minimum, and average number of data points for each WSD target word

For fair comparison with Tanabe (2020) in terms of the number of training data points, we conducted experiments without development data using CHJ-WLSP 2019. Here, the ratio of the training and test data was set to 4:1. The data split was performed using random sampling. We set epoch number as 20 and learning rate as 0.00005 according to the preliminary experiments. The results are an average of three trials.

In addition, for CHJ-WLSP 2019 and 2022, we used grid search for the hyper-parameters, i.e., the epoch number and learning rate using five-fold cross validation with development data. The options of the epoch number were from 1 to 30 and those of the learning rate were 0.00001, 0.0001, and 0.001. Here, the ratio of the training, test, and development data was set to 3:1:1. Please note that, when we performed cross validation, because we used a development set and Tanabe (2020) did not, the amount of training data of the BERT model is smaller than that of the model of the prior work.

Table 4 shows the number of training data points according to the experiment. Random sampling 2019 in the table means data split with random sampling for CHJ-WLSP 2019, and cross validation 2019 and 2022 indicate the cross validation using CHJ-WLSP 2019 and CHJ-WLSP 2022.

Experiments	Num of data
Random sampling 2019	134.26
Cross validation 2019	100.70
Cross validation 2022	1,353.09

Table 4 the number of test data points according to the experiment

1.1 Baseline

We used the method of Tanabe (2020) for the baseline. The best method in Tanabe (2020) is the method using fine-tuning contemporary features with a historical corpus in the Target Only scenario, i.e., the scenario where no example from contemporary corpus was used. She used NWJC2VEC, the word embeddings generated from the NWJC-2014-4Q dataset (Asahara et al., 2014) as pretrained word embeddings of contemporary Japanese texts. She used plain texts of CHJ to fine-tune word embeddings. She also used BCCWJ for fine-tuning and BCCWJ-WLPS for a sense-tagged contemporary corpus. BCCWJ is tokenized with contemporary UniDic (Den et al., 2010) and CHJ is tokenized with historical UniDic (Ogiso et al., 2012).

Tanabe (2020) used a scikit-learn library of support vector machine using NWJC2VEC fine-

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

tuned with CHJ. When generating the word embeddings of historical texts for a comparison, she used word2vec and the dimensionality was set to 200 and the window size was set to 2. The other parameters were the same as the default settings of the Gensim toolkit. She used five-fold cross validation without a development set. Please note that our MFS of the test data is slightly different from that of Tanabe (2020). We believe that is because the data split of cross validation is different.

Results

Table 5 shows the WSD accuracies of the models trained with CHJ-WLSP 2019. Hereinafter, Micro and Macro in the tables are micro- and macro averaged accuracies.

Model	Micro	Macro
Simple BERT model	**77.50%**	**72.82%**
Multitask learning	77.24%	72.55%
MFS of random sample	73.79%	69.81%
Tanabe (2020)	74.83%	70.80%
MFS of Tanabe (2020)	75.54%	70.00%

Table 5 WSD accuracies of the models trained with CHJ-WLSP 2019

This table shows that we substantially outperformed the MFS baseline. In addition, although our MFS of the test data is lower than that of the prior work (They are 75.54% and 73.79% when the micro-averaged MFS is compared), our BERT model significantly outperforms the result of the prior work according to the chi square test. The level of the significance in the test was 0.01. Additionally, the difference between our MFS and the BERT model was also significant.

In addition, Table 6 displays the WSD accuracies of the models trained with CHJ-WLSP 2019 using cross validation. Tables 5 and 6 indicate that multitask learning of WSD and document classification was not effective for CHJ-WLSP 2019. The differences between the simple BERT model and multitask learning were not significant. However, the two models using BERT outperformed the MFS baseline again.

Model	Micro	Macro
Simple BERT model	**76.85%**	**69.81%**
Multitask learning	76.70%	69.64%
Our MFS	74.20%	69.09%

Table 6 WSD accuracies of the models trained with CHJ-WLSP 2019 using cross validation

Table 7 shows the WSD accuracies of the models trained with CHJ-WLSP 2022 using cross validation. The table shows that multitask learning of WSD and document classification surpassed the simple BERT model. It was significant according to the chi square test. The level of the significance in the test was 0.01. In addition, the differences between the MFS and two BERT based models were also significant.

Model	Micro	Macro
Simple BERT model	84.68%	84.25%
Multitask learning	**85.17%**	**84.45%**
MFS	78.29%	78.20%

Table 7 WSD accuracies of the models trained with CHJ-WLSP 2022 using cross validation

Discussion

According to the three tables in Section 6, we can see that BERT trained from contemporary Japanese texts is effective for WSD of historical Japanese texts.

We believe that the reason why the BERT model outperformed the prior work that used word2vec is not the amount of training data that trained the BERT model or word2vec, because Tanabe (2020) used NWJC2VEC, the word embeddings generated from the NWJC-2014-4Q dataset, which included more than a billion sentences.

As Japanese BERT is trained with Japanese Wikipedia, which consists of approximately 17M sentences, the training data cannot be the reason. Therefore, even if we cannot know the concrete information that provided the improvement of the WSD performances, the network architecture could be the reason. In addition, as some studies reported BERT can capture various language information including syntactic structures (Jawahar et al. 2019), this property of BERT could be the reason of the success of the diachronic adaptation of historical Japanese.

We feared that the unknown words of the BERT model adversely affected the WSD accuracies, but they were at least not serious. Table 8 shows the percentage of unknown word tokens ([UNK]) and subword tokens that begins with a sharp mark (#) in the input tokens of the system. It means the tokens of input sentences are counted, excluding tokens beyond 512 tokens per an input. [UNK] and subword # in the table mean the percentage of unknown word tokens and that of the subword tokens that begin with a sharp mark, respectively.

Data	[UNK]	Subword #
CHJ-WLSP 2019	0.60%	11.71%
CHJ-WLSP 2022	1.52%	9.58%

Table 8 The percentage of unknown word tokens ([UNK]) and subword tokens that begin with a sharp mark (#)

According to Table 8, we can see that the unknown word tokens are very rare in CHJ-WLSP 2019 and CHJ-WLSP 2022 Because of subword tokens, most of input tokens could be interpreted by the BERT model trained with contemporary texts.

Next, let us discuss the multitask learning of WSD and document classification. When we used CHJ-WLSP 2019, this method did not work but when we used CHJ-WSLP 2022, it was significantly effective. The first reason to explain this fact that we can think of is the amount of the training data points, as shown in Table 4. The second reason could be the difference of the variety of the literature of two datasets. Tables 9 and 10 display the average number of test data in CHJ-WLSP 2019 and CHJ-WLSP 2022, respectively.

Literature	Num of test data
Taketori Monogatari	350.6
Tosa Nikki	264.8
Hōjōki	134.8
Tsurezuregusa	1,104
Toraakira-bon Kyogen	92.6

Table 9 Average number of test data in CHJ-WLSP 2019 disaggregated by the literature

Literature	Num of test data
Taketori Monogatari	312.2
Tosa Nikki	224.4
Hōjōki	120.6
Tsurezuregusa	1,011.2
Toraakira-bon Kyogen	22.8
Konjaku Monogatarishū	4,796
Uji Shūi Monogatari	3,167
Jikkinsyo	1,848.8
Taiyo Magazine	1,013.2
Kokutei Textbook	2,375.2

Table 10 Average number of test data in CHJ-WLSP 2022 disaggregated by the literature

According to Table 9, in CHJ-WLSP 2019, more than half of the test data came from only one literature, Tsurezuregusa. On the other hand, as shown in Table 10, the data balance is more balanced compared to CHJ-WLSP 2019.

Moreover, Tables 11 and 12 show the accuracies of document classification of CHJ-WLSP 2019 and CHJ-WLSP 2022. The improvement of the accuracy, that is the difference between accuracies of multitask learning and most frequent document in CHJ-WSLP 2022 was considerably higher than that in CHJ-WSLP 2019. This fact indicates that the document classification task was learned better using CHJ-WLSP 2022.

Model	Micro	Macro
Multitask learning	66.17%	63.95%
Most frequent document	58.15%	55.69%

Table 11 Accuracies of document classification of CHJ-WLSP 2019

Model	Micro	Macro
Multitask learning	63.90%	69.22%
Most frequent document	35.08%	36.53%

Table 12 Accuracies of document classification of CHJ-WLSP 2019

Finally, comparing the experiments using CHJ-WLSP 2019 and CHJ-WLSP 2022, we can see that the obvious factor to improve the WSD accuracies is the amount of in-domain labeled data. The amount of training data points of CHJ-WLSP 2022 was approximately 13 times more than that of CHJ-WLSP 2019 as shown in Table 4. This research showed that, when we use more than 1,000 data points for training data, the WSD accuracy is around 85%, which is considerably higher than the MFS baseline.

In the future, we plan to develop an all-words WSD system for historical Japanese texts. In addition, we plan to explore properties other than data amount to improve the performance of WSD.

Conclusions

We reported that BERT trained with contemporary Japanese texts considerably improved the WSD accuracies of historical Japanese texts. This method can be considered as a form of diachronic adaptation. Because the amount of training data of BERT model cannot account for the improvement of WSD accuracies, the network architecture of the BERT itself could be the reason why diachronic adaptation using BERT trained with contemporary text worked. In addition, because of the subword tokens, unknown word tokens are rare in historical texts. We performed experiments using two sets of a corpus, which are CHJ-WLSP 2019 and CHJ-WLSP 2022. We also showed the effectiveness of the multitask learning of WSD and classification of the sentence that was included the target word taken from, when we used CHJ-WLSP 2022. We also showed that the WSD accuracies substantially improved as the in-domain labeled data for training increased.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17H00917, 17KK0002, 18K11421 and 22K12145, and projects of the Center for Corpus Development, National Institute for Japanese Language and Linguistics.

References

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, iroya Takamura, and Daichi Mochihashi. 2021. Tsujiteki na tango no imihenka wo toraeru tango bunsanhyougen no ketsugogakusyu. [joint learning of word embeddings capturing diachronic changes of meanings of words]. In Proceedings of the NLP2021, (In Japanese), pages 712–717.

Masayuki Asahara, Nao Ikegami, Tai Suzuki, Taro Ichimura, Asuko Kondo, Sachi Kato, and Makoto Yamazaki. 2022. CHJ-WLSP: Annotation of `Word List by Semantic Principles' Labels for the Corpus of Historical Japanese, In proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, (To appear)

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiv ing and analysing techniques of the ultra-large-scale web-based corpus project of ninjal. Alexandria: The journal of national and international library and in formation issue, 25(1-2):129–148.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. Proceedings of ACL.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In Proceedings of ACL 2007, pages 256–263.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pages 23–59.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, Hideki Ogura. 2008. A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation, In Proceedings of the sixth international conference on Language Resources and Evaluation (LREC 2008), pages.1019-1024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2116–2121.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1501.

Sho Hoshino, Yusuke Miyao, Shunsuke Ohashi, Akiko Aizawa, and Hikaru Yokono. 2014. Machine translation from historical Japanese to contemporary Japanese using parallel corpus. In Proceedings of the NLP2014, (In Japanese), pages 816–819.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In Proceedings of ACL 2016, pages 897—-907.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the ACL 2019, pages 3651–3657.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 61–65.

Kazuma Kobayashi, Taichi Aida, and Mamoru Komachi. 2021. Bert wo shiyouu shita nihongo no tango no tsuujiteki na imihenka no bunseki. [analysis of diachronic changes in meanings of Japanese words using bert]. In Proceedings of the NLP2021, (In Japanese), pages 952–956.

Kanako Komiya and Manabu Okumura. 2011. Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning. In Proceedings of IJCNLP 2011, pages 1107–1115.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In Proceedings of the 24th International Conference on World Wide Web, pages 625–635.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 544 pages 3514–3520.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of Balanced Corpus of Contemporary Written Japanese. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), pages 1483–1486.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. Language resources and evaluation, 48(2):345–371.

Tatsuo Miyajima, Hisao Ishii, Seiya Abe, and Tai Suzuki. 2014. Nihon koten taisho bunrui goi hyo [Word list by semantic principles refereeing to Japanese classics]. Kasama Shoin, In Japanese.

National Institute for Japanese Language and Linguistics. 1964. Word List by Semantic Principles. Shuuei Shuppan, In Japanese.

Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den, and Yuji Matsumoto. 2012. Unidic for early middle Japanese: a dictionary for morphological analysis of classical Japanese. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 911–915.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. Semeval-2010 task: Japanese wsd. In Proceedings of the SemEval-2010, ACL2010, pages 69–74.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In Proceedings of EMNLP 2017, pages 1156–1167.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Semeval-2007 task 07: Coarse-grained english all-words task. In Proceedings of EACL 2017, pages 99–110.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017. Japanese all-words wsd system using the kyoto text analysis toolkit. In Proceedings of PACLIC 2017, pages 392–399.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. All words word sense disambiguation using concept embeddings. Proceedings of LREC 2018, pages 1006–1011.

Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, and Kanako Komiya. 2020. Neural machine translation from historical japanese to contemporary japanese using diachronically domain-adapted word embeddings. In Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 2020), page no. 22.

Aya Tanabe. 2020. Domain adaptation of word sense disambiguation of corpus of historical japanese using contemporary japanese corpus. Master theses of Graduate Schools of Science and Engineering, Ibaraki University of 2019 academic year (in Japanese).

Aya Tanabe, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. Detecting unknown word senses in contemporary japanese dictionary from corpus of historical japanese. In Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH 2018), pages 169–170.

Appendix A. Target Words

Word	Pronunciation	Translation	Year
為る	Suru	do	Both
一	Ichi	one	Both
居る	Iru	stay	Both
見る	Miru	look	Both
言う	Iu	say	Both
行く	Iku	go	Both
此れ	Kore	this	Both
今	Ima	now	Both

思う	Omou	think	Both
事	Koto	thing	Both
時	Toki	time	Both
取る	Toru	get	Both
所	Tokoro	place	Both
心	Kokoro	heart	Both
人	Hito	human	Both
成る	Naru	become	Both
知る	Shiru	know	Both
日	Hi	day	Both
物	Mono	object	Both
聞く	Kiku	listen	Both
又	Mata	and	Both
有る	Aru	there is	Both
様	Sama	appearance	Both
来る	Kuru	come	Both
或る	Aru	a certain	2019
下	Shita	under	2019
何	Nani	what	2019
家	Ie	house	2019
皆	Mina	every	2019
間	Aida	between	2019
共	Tomo	together	2019
月	Tsuki	moon	2019
見える	Mieru	see	2019
後	Ato	after	2019
国	Kuni	country	2019
作る	Tsukuru	make	2019
持つ	Motsu	hold	2019
書く	Kaku	write	2019
女	Onna	woman	2019
上	Ue	up	2019
身	Mi	body	2019
他	Hoka	other	2019
男	Otoko	man	2019
置く	Oku	put on	2019
中	Naka	inside	2019
道	Michi	way	2019
読む	Yomu	read	2019
内	Uchi	inside	2019
入る	Hairu	enter	2019
年	Toshi	year	2019
彼	Kare	he	2019
付ける	Tsukeru	put onand	2019
返る	Kaeru	return	2019
方	Hou	direction	2019
万	Man	ten thousand	2019
唯	Tada	only	2019
立つ	Tatsu	stand	2019
良い	Yoi	good	2019
我	Ware	I	2022
者	Mono	person	2022
出でる	Ideru	go out	2022
申す	Mousu	say	2022
是	Kore	this	2022
然る	Saru	like that	2022
其	Sore	that	2022
程	Hodo	level	2022
無い	Nai	no	2022

Table 2 Target words of WSD in CHJ-WLSP 2019 and CHJ-WLSP 2022

Gain-framed Buying or Loss-framed Selling?
The Analysis of Near Synonyms in Mandarin in Prospect Theory

Xin Luo
Dept of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
xin-tracy.luo@connect.polyu.hk

Chu-Ren Huang
Dept of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Abstract

While prospect theory (Kahneman and Tversky, 1979) has been widely used as a descriptive theory to explain various phenomena such as insurance and the relation between spending and saving, no studies have been found to look at prospect theory from linguistic prospective to investigate near synonymous verbs. This study investigates near-synonymous verbs in Mandarin Chinese under the framework of prospect theory.

Using the large-scale Gigaword2 Corpus, we examined how verbal transaction reflects gain and loss frame in terms of human decision. It is observed that mai3 'buy', mai4 'sell' and zu1 'rent' have default gain-loss frame encoded in the lexical meaning. Because of the topic-prominent typological feature of Mandarin, gain and loss frame have close relation with information structure. By using highlighting and omission of argument, the gain-loss frame can be emphasized. Our current study focuses on near synonymous verbs in Mandarin Chinese. It is hoped that more near synonymous verbs in Mandarin Chinese and in other languages can be investigated under the framework of prospect theory.

1 Introduction

According to Cruse (2022; 1986), near-synonyms can be defined as words that one or more of their senses have a sufficiently close similarity. In Chinese, mai3 'buy' and mai4 'sell' is one of the opposites pairs in Ding (2018)'s analysis of the large-scale corpus-based study on comparing the semantic use. In fact, extensive research have shown the semantic difference of the pair of 'buy' and 'sell' in English. They can be considered to be synonym or directional opposites because they can be expressed in two ways: A buys something from B or B sells something to A (Lyons, 1977). Cruse (2000) proposed that 'buy' and 'sell' may be four-place converses as the relational predicate they denote can have 4 arguments:

John sold the car to Bill for £5,000.
Bill bought the car from John for £5,000.

Mai3 'buy', mai4 'sell' and zu1 'rent', actually have the same eventive structure involving movement of money and merchandise. So far, however, no research has been found that investigate near synonymous verbs in the domain of gain-loss framing, especially based on natural language data in corpus.

Stemmed by Bernoulli (1954), expected utility is used as a criterion to explain the choice between gambles. This work was further developed into a descriptive theory of choice called *prospect theory* (Kahneman and Tversky, 1979) to evaluate the decision making under risk. Under this framework, it is observed that people tend to risk-averse with respect to gain and risk-seeking with respect to loss in different domains including financial decision and risky gambles. For example, the computerised laboratory experiments conducted by Powell and Ansic (1997) found that males are more risk preference whereas females are more risk-aversion in the financial decision making. From linguistics prospective, Zeng et al. (2022a) found that WAR metaphors were constantly sued with gain framing effects whereas loss-framed WAR metaphors were less frequently used. In terms of effectiveness of vaccination advocacy messages. Zeng et al. (2022b)'s experimental studies showed that gain-framed messages are more effective than loss-framed messages. To provide further investigation of prospect theory from linguistic prospective, the current study will attempt to examine whether near-synonymous verbs mai3 'buy', mai4 'sell' and zu1 'rent' can reflect the gain and loss frame and how the gain and loss can be reflected on lexical choice and syntactic structure in decision-making problems.

2 Literature review

2.1 Prospect theory

Prospect theory (Kahneman and Tversky, 1979) is an alternative model of expected utility theory (von Neumann et al., 1944) to describe individual decision making under risk. This classic work with enormous influence and has been used as a descriptive theory to explain various phenomena such as insurance and the relation between spending and saving.

The gain and loss in prospect theory is measured by the outcomes. In other words, whether buying and selling are expressed positive or negative deviations (gains or losses) depends on the evaluation towards a reference point (neutral reference outcome). Notably, the variation of the reference point can determine whether an outcome is evaluated as a gain (positive) or a loss (negative) as shown in the S-shaped value function in Figure 1. An example given by Tversky and Kahneman (1981) is the cost of the purchase of a car. The transaction of the whole purchase is evaluated as positive, negative or neutral depending on various factors such as the performance of the car and the price of similar cars in the market.

Under the framework of prospect theory, Mather et al. (2012)'s experimental study examined the relationship of risk preferences and aging in domains of certain and risky gambles and Hameleers (2021) assessed the effect of gain-loss framing on risky choices and emotional response in times of the pandemic. While many of the experimental studies enrich the analysis in this field from different angles, the experiments are mainly based on survey and pre-designed tasks but rarely based on natural linguistic data. Furthermore, many of the current studies are conducted in western countries but not in China. Several others are in health domain (Kim, 2012) but not in decision-making problems.

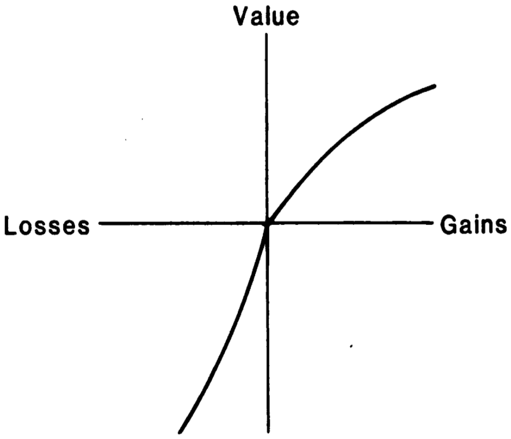


Figure 1: A hypothetical value function.

From the linguistic perspective, there is default gain-loss frame encoded in the language which is similar to the valence of emotion words. For example, 喜歡 'like' is the positive emotion word whereas 愁 'worry' is the negative emotion word. Mai3 generally refers to the acquiring of something, it is encoded in a gain frame from linguistic perspective. For mai4 and zu1, they are directional because they involve giving away something and acquiring the money at the same time. They, therefore, can be in either gain or loss frame. When individual choose to make decision of buying rather than selling, the whole transaction is encoded in the default gain frame.

2.2 Thematic structure in Mandarin Chinese

Thematic structure which views the clause as message is made up of two distinct parts, *Theme* and *Rheme*. According to Halliday and Matthiessen (2014), the thematic structure is expressed by a Theme accompanied by a Rheme. The Theme appear in the sentence-initial position which serves as 'the point of departure of the message' and followed by Rheme, 'the part in which the Theme is developed'. The sentence-initial Theme highlights the thematic prominence information for the addressee (Halliday and Matthiessen, 2014). Some scholars have used the term *Topic* and *Comment* instead of Theme and Rheme (e.g., Li and Thompson, 1981). In this study, we follow Li and Thompson (1981) and retain the terminology of Topic and Comment.

In Mandarin, topic-prominent sentence structure is a significant typological feature of Mandarin. In the topic-prominent sentence structure, word order is governed by meaning rather than grammatical function. This means that sentence with omitted subject, sentence with object at the beginning can be found in Mandarin. Topic, information known to both the speaker and the hearer, occurs in the preverbal position whereas comment, the new information occurs in the postverbal position. Take the short narrative from (Shyu, 2016) as an example.

[1]
a. 我特別喜歡讀...賈平凹的書。
wo3 te4bie2 xi3huan1 du2 jia3ping2wa4 de0 shu1
'I especially like to read books by JIA Pingwa.'

b. 這兩本書我都看爛了。
zhe4 liang3 ben3 shu1 wo3 dou1yao4 kan4 lan4 le0
'These two books are worn and torn because I have read them so often.'

The first sentence [1a] is in canonical sentential order of SVO. The underlined post-veral object, books written by JIA Pingwa, is the new information and also the focus of the sentence. When additional information is added to the two books in [1b], the definite nominal 這兩本書 'these two books' becomes the shared/old information at the sentence-initial position with remaining comment clause to elaborate on.

3 Methodology and research questions

This study adopts a corpus-based approach to investigate the usage of three transitive verbs mai3 'buy', mai4 'sell' and zu1 'rent' from a linguistic point of view under the framework of prospect theory. The data are extracted from the large-scale corpus of Gigaword2 in Chinese Word Sketch (CWS) (Huang, 2009) which is one of the largest balanced corpora of collected news texts from Taiwan, Mainland China and Singapore. As a *topic-prominent* language, the 'topic' which is the given information usually occur in the pre-verbal position and followed by new information in Mandarin Chinese. Based on the topic-prominent feature in Mandarin Chinese, we examined how verbal transaction reflect gain and loss frame in terms of human decision.

4 Findings and Discussion

The following sessions present the findings and discussion of the results. From the search results in Chinese WordNet in Figure 2, these transitive verbs have the same eventive structure but in different directions. The default gain-loss frame encoded in the verbal transaction can be observed in general definition of the words. The linguistic encoding of mai3 'buy' is a default gain frame because people gain the merchandise. However, mai4 'sell' and zu1 'rent' are directional because they involve movement of money and merchandise. Therefore, mai4 'sell' and zu1 'rent' can be in gain or loss frame.

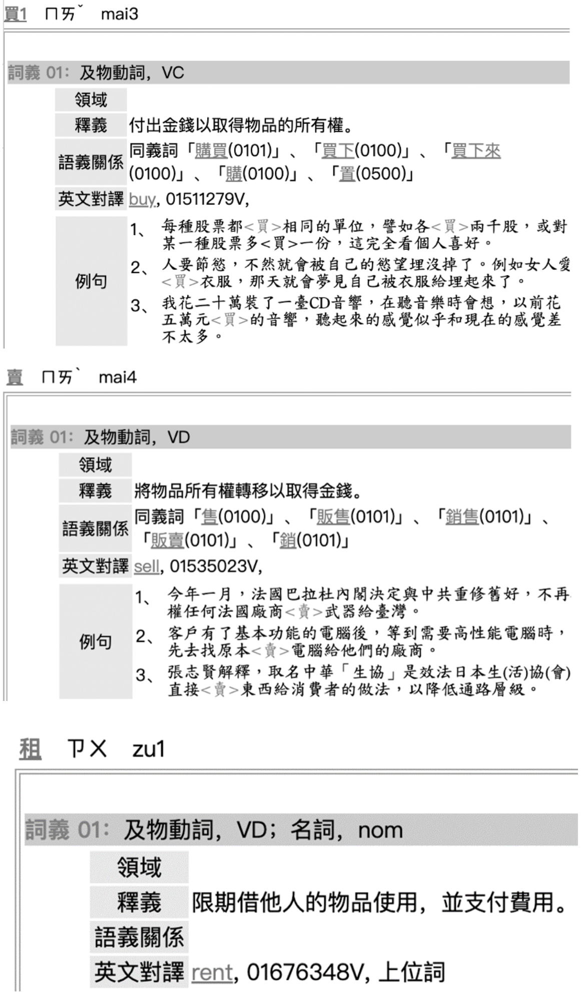
Table 1 shows the PoS of mai3 'buy', mai4 'sell' and zu1 'rent' and their frequency in Gigaword 2 Corpus. Mai3 'buy' has the highest frequency of 45,553 in Gigaword 2, followed by mai4 'sell' and zu1 'rent'. In terms of part of speech, mai3 'buy' is a transitive verb which has an object to take the argument role whereas mai4 'sell' and zu1 'rent' are ditransitive verbs which have two objects to take the argument role.

Verb	PoS	Frequency
mai3 'buy'	VC31	45,553
mai4 'sell'	VD1	29,494
zu1 'rent'	VD	5,938

Table 1: PoS of mai3 'buy', mai4 'sell' and zu1 'rent'

4.1 Gain-loss frame encoded in buying

Table 2 shows the subject and object of mai3 'buy' with the top five MI values. Mai3 'buy' is a two-argument verb that takes a subject (the one that pay) and an object (the merchandise to be gained). It is found that the subject of mai3 is usually the person to pay such as 顧客 'customer', 我 'I', 你 'You' and 自己 'self'. In addition, 錢 'money' can also be the subject. The merchandise to be gained such as 房子 'house', 東西 'thing' are usually take the object position.



買1 ㄇㄞˇ mai3

詞義 01: 及物動詞，VC

領域	
釋義	付出金錢以取得物品的所有權。
語義關係	同義詞「購買(0101)」、「買下(0100)」、「買下來(0100)」、「購(0100)」、「置(0500)」
英文對譯	buy, 01511279V,
例句	1、 每種股票都<買>相同的單位，譬如各<買>兩千股，或對某一種股票多<買>一份，這完全看個人喜好。 2、 人要節慾，不然就會被自己的慾望埋沒掉了。例如女人愛<買>衣服，那天就會夢見自己被衣服給埋起來了。 3、 我花二十萬裝了一臺CD音響，在聽音樂時會想，以前花五萬元<買>的音響，聽起來的感覺似乎和現在的感覺差不太多。

賣 ㄇㄞˋ mai4

詞義 01: 及物動詞，VD

領域	
釋義	將物品所有權轉移以取得金錢。
語義關係	同義詞「售(0100)」、「販售(0101)」、「銷售(0101)」、「販賣(0101)」、「銷(0101)」
英文對譯	sell, 01535023V,
例句	1、 今年一月，法國巴拉杜內閣決定與中共重修舊好，不再權任何法國廠商<賣>武器給臺灣。 2、 客戶有了基本功能的電腦後，等到需要高性能電腦時，先去找原本<賣>電腦給他們的廠商。 3、 張志賢解釋，取名中華「生協」是效法日本生(活)協(會)直接<賣>東西給消費者的做法，以降低通路層級。

租 ㄗㄨ zu1

詞義 01: 及物動詞，VD；名詞，nom

領域	
釋義	限期借他人的物品使用，並支付費用。
語義關係	
英文對譯	rent, 01676348V, 上位詞

Figure 2: Search result of mai3 'buy',mai4 'sell' and zu1 'rent' in Chinese WordNet.

However, because of the topic-prominent sentence structure in Mandarin Chinese, the Topic of a sentence is placed preverbally. The Topic is the given information which the speaker and listener already known, as in the following examples.

[2]
a. 房子買在山東。
fang2zi mai3 zai4 shan1 dong1。
'The house was bought in Shandong.'

b. 他在温哥華買了一棟房子。
ta1 zai4 wen1 ge1 hua2 mai3 le0 yi1 dong4 fang2zi0
He bought a house in Vancouver.

c. 兩萬塊錢買個烏紗帽。
liang3 wan4 kuai4 qian2 mai3 ge4 wu1 sha1 mao4
'The black gauze cap costs 20 thousand dollars.'

Subject of mai3	Frequency	MI value
錢(money)	635	56.7
顧客(customer)	103	37.8
我(I)	237	34.1
你(you)	111	33.5
自己(myself)	290	32.4
Object of mai3	**Frequency**	**MI value**
房子(house)	864	65.4
東西(thing)	1280	64.7
書(book)	811	54.8
衣服(cloth)	335	49.1
房(house)	219	47.1

Table 2: Most frequent subject and object for mai3 'buy'.

房子 'house' in [2a] is the given information because it is the person known to speaker and listener. When additional information is added to describe who purchases the house, 房子'house' becomes the new information in [2b]. When the numeral-measure phrase occurs in the preverbal position, it signals its status that speaker and listener already know the amount of money, the more informative is 烏紗帽 'the black gauze cap' obtained with this amount toward the end of the sentence as in [2c].

It is observed that the monetary payment can be placed in between the subject and the verb mai3 'buy'. The monetary payment can give additional neutralized information, as in [3a] and [3b] or mitigate the gain frame of the sentence, as in [3c-3e].

[3]
a. 老百姓掏錢買房子。
lao3bai3xing4 tao1 qian2 mai3 fang2zi0。
Ordinary people spend money to buy houses.

b. 15 歲時勞達向奶奶借錢 買了一輛大衆牌二手車。
15 sui4 shi2 lao2 da2 xiang4 nai3nai0 jie4 qian2 mai3 le0 yi1 liang4 da4 zhong4 pai2 er4shou3 che1
'When Lauda was 15, he borrowed money from his grandmother to buy a second-handed Volkswagen.'

c. 家長抱怨其子女花費太多錢 買耶誕節卡片

和禮物。

jia1chang2 bao4yuan4 qi2 zi3nv3 hua1fei4 tai4duo1 qian2 mai3 ye2dan4jie2 ka3pian4 he2 li3wu4。

'Parents complain that their children spend too much money on Christmas cards and presents.'

d. 這車子是貸款 買的，車主還是車行。

zhe4 che1 zi0 shi4 dai4kuan3 mai3 de0，*che1zhu3 hai2 shi4 che1hang2*

'The car was bought on loan. The owner is still the car dealer.'

e. 野村證券高價 買股票彌補主要客戶的損失。

ye3cun1 zheng4quan4 gao1 jia4 mai3 gu3piao4 mi2bu3 zhu3yao4 ke4hu4 de0 sun3shi1

'Nomura bought shares at high prices to cover the losses of key clients.'

In [3a] and [3b], 掏錢 'spend money' is in a preverbal position, which signifies the relation between the verb and the nominal phrase. Likewise, 向奶奶借錢 'borrow money from grandma' provides additional information of the monetary exchange. In contrast, the preverbal monetary payment can emphasize certain loss. 花費太多錢 'spend too much money' in [3c], 貸款 'on loan' in [3d] and 高價 'bought at high prices' in [3e] have weaken the gain frame by highlighting certain loss. Furthermore, the verb of mental activity 抱怨 'complain' in [3c] further confirm the mitigation of gain frame. Likewise, 車主還是車行 ' The owner is still the car dealer' in [3e] and 彌補損失 'to cover the losses' are new information which further strengthen the mitigation of gain frame.

From the above observation, we found that mai3 has a predominant structure of gain frame. However, the gain frame can be mitigated or weakened by certain loss of money or compensation which occur in the preverbal position.

4.2 Gain-loss frame encoded in selling

Mai4 is a ditransitive verb which involves one subject (an agent who initiates the transfer) and two objects (a theme that is transferred and a beneficiary or maleficiary who receives or loses the theme).

[4] 任何人不得賣酒給二十一歲以下的年輕人。

ren4he2 ren2 bu4de2 mai4 jiu3 gei3 er4shi2yi1 sui4 yi3xia4 de0 nian2 qing1 ren2

'No one can sell alcohol to young people under twenty-one.'

Subject of mai4	Frequency	MI value
門票	38	29.9
菜	28	29.1
專輯	29	25.4
農副產品	23	24.8
低價	14	19.8
Object of mai4	**Frequency**	**MI value**
菜	607	69.3
價錢	299	61.2
錢	395	40.3
肉	112	40.2
假貨	60	38.7

Table 3: Most frequent subject and object for mai4 'sell'.

In example [4], 任何人 'anyone' is the agent who initiates the wine selling. Two objects in this example are young people (goal) and the wine as the theme. Table 3 shows the subject and object of mai4 'sell' of top five MI values. Different from mai3 'buy', the nouns which take the subject position for mai4 'sell' are usually the products to be sold whereas the object would be the price or the products.

Mai4 'sell' usually involves double movement of money and merchandise, as shown in [5a]. However, the argument can be omitted to change the focus of the sentence. Compare the following examples.

[5]

a. 這位44歲的彝族壯漢和他的妻子一道靠賣土豆賺了不少錢 。

zhe4 wei4 44 sui4 de0 yi2zu2 zhuang4 han4 he2 ta1 de0 qi1zi0 yi1 dao4 kao4 mai4 tu3dou4 zhuan4 le0 bu4shao3 qian2

'The 44-year-old man from Yi ethnic group makes a lot of money by selling potatoes with his wife.'

b.這位44歲的彝族壯漢和他的妻子一道賣土豆。

zhe4 wei4 ４４*sui4 de0 yi2zu2 zhuang4 han4 he2* *ta1 de0 qi1zi0 yi1 dao4 mai4 tu3dou4*

'The 44-year-old man from Yi ethnic group sells potatoes with his wife.'

c. 這位44 歲的彝族壯漢和他的妻子一道賺了不少錢 。

zhe4 wei4 44 sui4 de0 yi2zu2 zhuang4 han4 he2 ta1 de0 qi1zi0 yi1 dao4 zhuan4 le0 bu4shao3 qian2

'The 44-year-old man from Yi ethnic group and his wife make a lot of money.'

In these examples, the given information is the same while the emphasis of the sentences is different. In [5b], the sentence focuses on the merchandise to be sold but omit the money they made. The loss frame is emphasized. In contrast, the products that being transferred can be omitted by highlighting the money obtained, as in [5c]. The gain frame is emphasized. Note that using omission of argument (the object being transferred or the money), the gain or loss frame for mai4 'sell' can be emphasized. Similarly, the gain frame of getting a good price is highlighted by omitting the argument of what to be transferred in [6]

[6] 三兩年内保證能賣上好價錢。

san1 liang3 nian2 na4 bao3zheng4 neng2 mai4 shang4 hao3 jia4qian2

'Guaranteed a good price in a couple of years.'

In addition, the patient of the action 黎明粵語專輯 'Leon's album' occupies the subject position while the agent phrase does not appear. The *topic-comment* sentences are commonly used in Mandarin Chinese to package the flow of information. When 黎明粵語專輯 'Leon's album' occurs at the preverbal position as the topic, it represents the speaker and listener wants to discuss again as a piece of the given information. The new information is the attractiveness of Leon's album. The *topic-comment* sentence in [8] is similar to that in [7] in that the 肉類 'meat' appears in the preverbal position to represent a piece of old information while the focus of the sentence is 好賣 'sells well'. Mai4 'sell' becomes intransitive and the price will not be mentioned.

[7] 黎明粵語專輯在香港大賣。

li2ming2 yue4 yu3 zhuan1ji2 zai4 xiang1gang3 da4 mai4

'Leon's Cantonese album sells well in Hong Kong.'

[8] 肉類依然好賣。

rou4 lei4 yi1ran2 hao3 mai4

'Meat still sells well.'

4.3 Gain-loss frame encoded in renting

Subject of zu1	**Frequency**	**MI value**
土地	58	29.4
國有地	5	18.4
水	11	14.3
人士	5	4.1
政府	10	3.7
Object of zu1	**Frequency**	**MI value**
住處	337	65.7
船	75	35.8
公寓	41	31.6
場地	50	29.6
攤位	32	29.1

Table 4: Most frequent subject and object for zu1 'rent'.

In section 4.1 and 4.2, we have investigated the gain-loss frame in association with mai3 'buy' and mai4 'sell'. Similar to mai4 'sell', zu1 'rent' is a ditransitive verb which involves two objects to take the argument role. However, zu1 is directional in either being a beneficiary to receive the theme or maleficiary who loses the theme. Table 4 shows the most frequent subject and object for zu1 'rent' in Gigaword2 Corpus.

[9]

a. 這些衛星也可部分地「租」給民用了。

zhe4 xie1 wei4xing1 ye3 ke3 bu4fen de0 「zu1 」gei3 min2yong4 le0

'These satellites can also be partly "leased" to civilian use.'

b. 長榮航空向英國MONARCH 租得的兩架波音七六七。

chang2 rong2 hang2kong1 xiang4 ying1guo2 MONARCH zu1 de2 de0 liang3 jia4 bo1yin1 qi1liu4qi1

'Eva Air leases two Boeing 767s from Britain's MONARCH.'

c.自己租了一輛大客車。

zi4ji3 zu1 le0 yi1 liang4 da4ke4 che1
'I rented a bus.'

d. 自己花了100塊租了一輛大客車。
zi4ji3 hua1 le0 100 kuai4 zu1 le0 yi1 liang4 da4ke4 che1
'I rented a bus for 100 yuan.'

In [9a], the beneficiary of zu1 'rent' is introduced by 給 'give' which emphasizes the loss frame. In [9b], the source is introduced by 向 'to'. It is the gain frame marked by 向...租得 'rent...to'. 長榮航空 'EVA Air' is the beneficiary who gains the money. In [9c], 一輛大客車 'a bus' occurs at the postverbal position, gain frame is thus emphasized. In [9d], the gain frame for zu1 can be mitigated by certain loss of 100 yuan.

5 Conclusion

Across the corpus-based study we look at prospect theory from a linguistic perspective to investigate the three near-synonymous verbs. We found that verbal transaction can reflect gain and loss frame in terms of human decision. It is observed that mai3 'buy', mai4 'sell' and zu1 'rent' have a default gain-loss frame encoded in the lexical meaning. Specifically, it is observed that mai3 has a predominant gain frame whereas mai4 and zu1 can be in gain or loss frame.

In addition, it is found that gain and loss frame have close relation with information structure because of the topic-prominent characteristic in Mandarin Chinese. By using highlighting and omission of argument, the gain-loss frame can be emphasized. Our current study focuses on near synonymous verbs, we hope the current study can be extended to study more near synonymous verbs in Mandarin Chinese and near synonymous verbs in other languages.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful and helpful comments.

References

Daniel Bernoulli. 1954. Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

D. A. Cruse, 2000. *Meaning in Language : An Introduction to Semantics and Pragmatics.*, chapter 9, pages 163–176. Oxford Textbooks in Linguistics. OUP Oxford.

D. A. Cruse. 2022. *A Glossary of Semantics and Pragmatics*. Edinburgh University Press.

Jing Ding. 2018. *A Lexical Semantic Study of Chinese Opposites*. Springer Singapore.

Michael A.K. Halliday and Christian Matthiessen, 2014. *Clause as Message*, volume 1, page 88–133. Routledge.

Michael Hameleers. 2021. Prospect theory in times of a pandemic: The effects of gain versus loss framing on risky choices and emotional responses during the 2020 coronavirus outbreak–evidence from the us and the netherlands. *Mass Communication and Society*, 24(4):479–499.

Chu-Ren Huang. 2009. Tagged chinese gigaword version 2.0.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Hyo Jung Kim. 2012. The effects of gender and gain versus loss frame on processing breast cancer screening messages. *Communication Research*, 39(3):385–412.

Charles N. Li and Sandra A. Thompson. 1981. *Mandarin Chinese: a Functional Reference Grammar*. University of California Press, Berkeley.

John Lyons, 1977. *Structural semantics II: sense relations*, volume 1, page 270–335. Cambridge University Press.

Mara Mather, Nina Mazar, Marissa A Gorlick, Nichole R Lighthall, Jessica Burgeno, Andrej Schoeke, and Dan Ariely. 2012. Risk preferences and aging: The "certainty effect" in older adults' decision making. *Psychology and aging*, 27(4):801.

Shu-Ing Shyu, 2016. *Information structure*, page 518–576. Reference Grammars. Cambridge University Press.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

John von Neumann, Oskar Morgenstern, Harold W. Kuhn, and Ariel Rubinstein, 1944. *Formulation of the Economic Problem*, pages 1–45. Princeton University Press.

Huiheng Winnie Zeng, Chu-Ren Huang, and Kathleen Ahrens. 2022a. Fighting against the pandemic: A gain-framed or loss-framed war? *Presented at the 30th Joint Workshop on Linguistics and Language Processing (JWLLP 2022)*, University of Macau, 26 March 2022.

Huiheng Winnie Zeng, Yin Zhong, Kathleen Ahrens, and Chu-Ren Huang. 2022b. Effects of gain/loss framing on pandemic vaccination responses. *To be presented at the 15th conference of the Association for Researching and Applying Metaphor (RaAM 15)*, University of Bialystok, Poland, 21-24 September 2022.

GPT-2 Contextual Data Augmentation for Word Sense Disambiguation

Rakia Saidi
LIMTIC Laboratory,
Faculty of science of Tunisia
UTM University / Tunisia

Fethi Jarray
LIMTIC Laboratory
ISI Medenine
Gabes University

Jeongwoo Jay Kang
LIG Laboratory, Emvista
Univ. Grenoble Alpes
France

Didier Schwab
LIG Laboratory
Univ. Grenoble Alpes
France

{saidi.rakya,fjarray}@gmail.com
{didier.schwab,jeongwoo.kang}@univ-grenoble-alpes.fr

Abstract

Most Word-Sense Disambiguation (WSD) systems rely on machine learning approaches that require large-scale corpora for effective training. So, the quality of a WSD system degrades when trained in a low-resource language such as Arabic. To improve WSD, we design a novel data augmentation technique by properly fine-tuning GPT-2 to generate a gloss or a phrase for a selected word. The generated training data is then combined with the original dataset to train a BERT-based WSD classifier. Experimental results show that integrating this augmentation technique improves WSD quality for both low-resource language (Arabic) and high-resource languages (English).

1 Introduction

Polysemous is a word with several related meanings. Some words, such as run or set have more than thirty different meanings. Polysemous words cause ambiguity in contexts where the meaning is different from the primary meaning of the word. For example, more than 40% of English words have more than one meaning (Nagy, 1995). In natural language processing, word sense disambiguation (WSD) consists of identifying the intended meaning (sense) of a polysemous word in a given context.

Two main approaches to WSD can be distinguished: Knowledge-Based Approach and Machine Learning-Based Approach. The former relies on external lexical resources, such as the well-known WordNet Knowledge Base. Machine learning based approaches use sense-annotated corpora to train the WSD system. The success of ML approaches in general and neural networks in particular highly depends on the availability of two resources inputs: 1) a dictionary or a sense inventory such as WordNet to specify senses for each word from some lexicon, 2) a sense annotated corpus for training such as SemCor. Deep neural networks achieved state-of-the-art performance for WSD. However, they are prone to overfitting on small datasets because they learn millions or billions of parameters while building the model. To avoid overfitting, the best solution is to use more training data by augmenting existing data.

WSD tasks suffer from insufficient training data or unavailability of annotated corpora. To overcome this challenge, we propose to generate new corpora or increase the size of existing ones. More specifically, we propose data augmentation techniques based on generative language models to improve the accuracy of learning methods.

In this paper, we start by discussing data augmentation (DA) in general and data augmentation for WSD in particular, with a focus on GPT(Generative Pre-trained Transformer) such as a generative language models. To validate our proposed method, we carried out experiments on GlossBert for English WSD and on ArabicGlossBert for Arabic. By conducting experiments on various different WSD tasks, we show that the proposed GPT data augmentation performs better than the baselines, existing techniques that are mainly devoted to the machine translation task.

The main contributions are as follows.

- We propose a GPT2 based augmentation method. The method allows GPT2 to augment sentences without altering the target word meaning. Our approach can further be applied to other NLP tasks such as semantic matching and natural language inference.
- Experimental results show that our approach achieves better performance, comparing with existing data augmentation methods.

Our contribution falls into centralized DL techniques, contrary to federated learning (Boughorbel et al., 2019). We suppose that the dataset set is clean, as opposed to the noisy data set (Boughorbel et al., 2018). To our knowledge, this is the first attempt to utilize GPT-2 to augment data for WSD tasks.

The remainder of this paper is structured as follows. Section 2 presents the related work on WSD-DA. Section 3 explains our DA method for WSD. Section 4 presents our integrated WSD system, where BERT models are used for feature extraction. Section 5 discusses the results obtained. We conclude this paper with a summary of our contribution and we mention some future extensions.

2 Related work

Data augmentation strategy is used in computer Vision and Natural Language Processing(NLP) to overcome the challenge of deal data scarcity and data diversity insufficiency and to prevent overfitting. Data augmentation in NLP is not sufficient to replace a word with its synonym because the context will be different and many downstream tasks depending on the context, so it is not as easy as in computer vision where image cropping showed good improvements,

Data augmentation is basically performed based on human knowledge of invariance, rules, or heuristics. According to Li et al. (2022), data augmentation techniques can be categorized into three main categories: paraphrasing, noising and sampling. Methods based on paraphrasing replace a word by a generated one with its synonyms (Zhang et al., 2015) or with the most similar word after a similarity computation(Wang and Yang, 2015). This approach can't ensure a good diversity due to the limitation of lexical ontology such as WordNet and thus the diversity. Noising-based approaches add discrete or continuous noise to expand the training set, such as randomly swapping two words or replacing a word with another(Xie et al., 2017). Even if noising is a popular technique in computer vision, it may alter the semantics of the sentence and will change the context. For example, in sentiment analysis, changing the word happy by sad will change the class of sentences.

To address the limitations of the aforementioned approaches, sampling methods based on contextual augmentation have been proposed. Fadaee et al. (Fadaee et al., 2017) focused on parallel corpora and replaced high-frequency words with rare words in the target language according to a language model, then changed source words accordingly. Kobayashi (Kobayashi, 2018) and Wu et al. (Wu et al., 2019) applied a language model on a target word and predict a new word that will replace the original one. Gao et al. (Gao et al., 2019) introduced a soft contextual augmentation that occurs at the word embedding level, where the target word embedding is replaced by the expectation of word embeddings of predicted words by a language model. Yand et al. (Yang et al., 2019) fine-tuned the BERT model on different data sets for Open-Domain Question Answering and considered it also as a data augmentation. Papanikolaou and Pierleoni(Papanikolaou and Pierleoni, 2020) proposed a GPT-2 based approach to augment data training in relation extraction task. Kun et al. (Li et al., 2020), proposed a conditional augmentation method based on sequence to sequence generation for aspect term extraction task.

The contextual augmentation preserves the semantic since the new word is generated from the entire context, not only from the current target as it done by the other approach. Nevertheless, none of the above methods is applicable for WSD, as they don't take into account the ambiguous target word. There are few works dedicated to WSD DA. Yap et al. (2020) noted that the majority of the synsets in WordNet contain illustrative short sentences. He included illustrative sentence-gloss on the set of context-gloss pairings. Using this technique, they obtained 37,596 additional training instances (about 17% more training instances).

Kohli (2021) adopted a back-translation strategy for data augmentation. They fine-tuned the GlossBERT model on the augmented corpus and evaluated it on semEval. The disadvantage of this technique is that some information may be lost in back-translating process.

Lin and Giambi (2021) used BERT and WordNet to investigate alternative data augmentation approaches on context-gloss pairs to improve the performance of WSD. They demonstrated that augmentation procedures at the sentence and word levels are effective solutions for WSD. They also discovered that adding hypernym glosses from a lexical knowledge base can increase performance. Their procedure is available only for the French, German and Russian languages (not for Arabic and English).

Yuan et al. (2016) constructed a semi-supervised system for WSD data augmentation. It consists of increasing the tagged sample sentences with a huge number of unlabeled sentences from the Web to solve these shortcomings by using the LSTM model. It is possible to add many unlabeled sentences by adding many unlabeled sentences. The data sets used in this work are SemEval and Semcor based on WordNet. The major drawback of this approach is that the generated sentences depend on the seed set.

Our work falls into the sampling techniques for text, but we focus on word-sense disambiguation.

3 Proposed method

To the best of our knowledge, this is the first contribution dedicated to data augmentation in WSD by the GPT family. GPT (Generative Pre-Training) is a large transformer-based language model with billions of parameters and trained on gigabytes of text scraped off the Internet. GPT-2 has shown impressive text generation results and low perplexity on several benchmarks. GPT-2 was trained with a standard language modeling objective and is therefore powerful in predicting the next word in a sequence of seen words. There are other variants of GPT that typically differ according to the number of layers, the size of the layers and the corpus used to train them such as GPT3, GPT-Neo and GPT-NeoX but we can't use it due to the huge size of these models and the unavailability of GPT-3.

In this paper, we aim to use GPT-2 to automatically generate WSD training data and combine it with existing datasets to train better WSD models. More concretely, given a training dataset for WSD such as Semcor, we fine-tuned the pre-trained GPT-2 model on this dataset to encourage GPT-2 to produce synthetic sentences that preserve the meaning of a target word.

The main challenge is how to deal with the target word and under which hypotheses, it retains its meaning. Our data augmentation is shown in Algorithm 3. Given an original tagged sentence, our strategy is to freeze the target word and to generate a gloss or phrase for a context word, such as the last word of a sentence. The number of context words replaced by a gloss can be seen as an extra hyperparameter.

[H]

Set a generation strategy

each sentence with an annotated target

- Freeze the target word
- Select context words according to the generation strategy
- Generate a gloss or a phrase for the selected word

We studied three strategies for increasing the number of examples based on a context selection of a chosen sense.

1. Entire context (Selection of the whole context): we replace each context word with a sequence of words. In this strategy, the sentence loses its structure;

2. Context-tail (Selection of the end of the context): we replace the last word of the context with a sequence of words.

3. Context-head (Selection of the beginning of the context): we replace the first word of the context with a sequence of words.

Figure 1 shows an example for each strategy of context selection. We note that we can exhibit other selection strategies, such as the random selection of a set of context words.

We validate our system on low-resource language (Arabic) and high-resource languages (English). So, we used a GPT-2 and a BERT models different for each language.

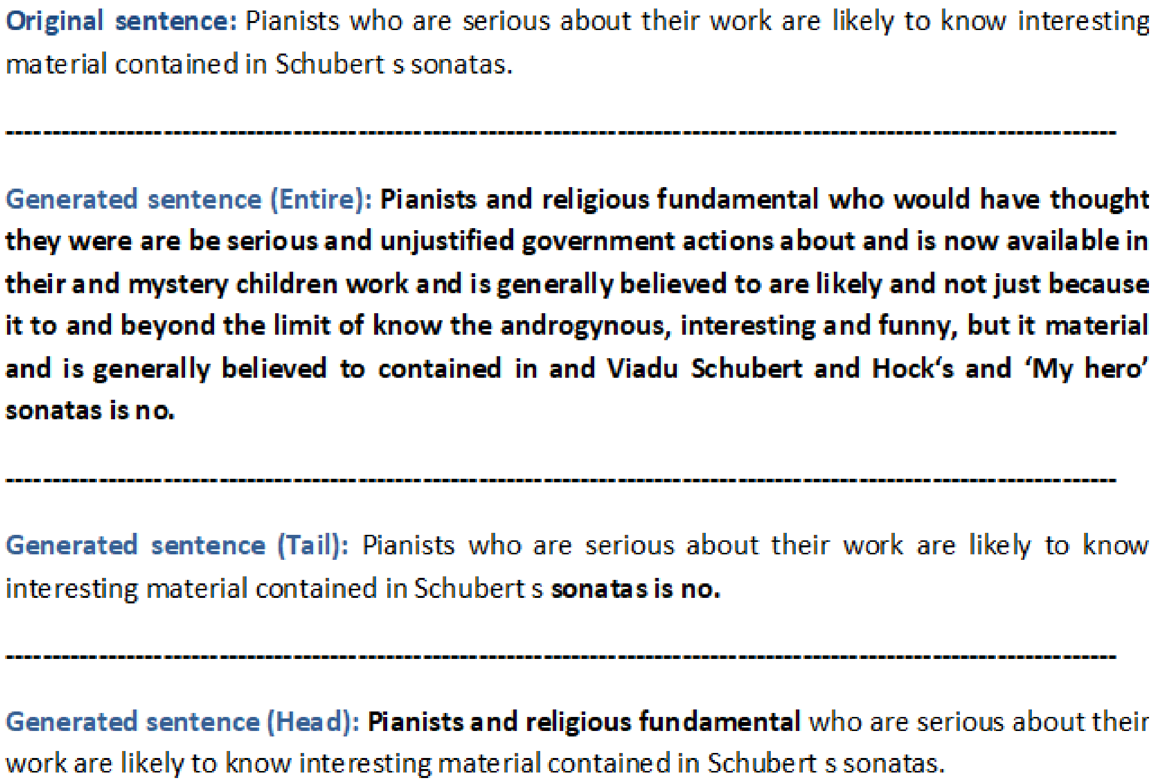
4 Experimental Evaluation

In this contribution, we are mainly interested in data augmentation for the task of lexical disambiguation for the Arabic and English languages. We use UFSAC (Vial et al., 2018) (*Unification of Sense Annotated Corpora and Tools*) which brings together the annotated corpus from the WordNet that have been automatically ported in Arabic and French via lexical transfer (Salah et al., 2018a).

4.1 Setup

To fine-tune GPT-2, we employed the medium model (355M) with a learning rate of 10^{-5}, a restore_from of fresh and a batch size of 16. The fine-tuning objective was to minimize the binary cross-entropy loss between the predicted senses and the golden senses. We applied the three above-mentioned strategies for Arabic and the English because it is possible to select any number of context words and replace them by a gloss.

We used pre-trained BERT models as a WSD classifier which we fine-tuned on either the gold or the gold+generated datasets. For fine-tuning, we used the pre-trained uncased BERT-BASE model. The total number of parameters in this pre-trained



Original sentence: Pianists who are serious about their work are likely to know interesting material contained in Schubert s sonatas.

Generated sentence (Entire): **Pianists and religious fundamental who would have thought they were are be serious and unjustified government actions about and is now available in their and mystery children work and is generally believed to are likely and not just because it to and beyond the limit of know the androgynous, interesting and funny, but it material and is generally believed to contained in and Viadu Schubert and Hock's and 'My hero' sonatas is no.**

Generated sentence (Tail): Pianists who are serious about their work are likely to know interesting material contained in Schubert s **sonatas is no.**

Generated sentence (Head): **Pianists and religious fundamental** who are serious about their work are likely to know interesting material contained in Schubert s sonatas.

Figure 1: Examples for each of the selection strategies (in order, Entire-context, Context-tail, Context-head)

model is 110M, with 12 Transformer blocks, 768 hidden layer blocks, and 12 self-attention heads. For the optimizer, we used Adam (Kingma and Ba, 2014), a sequence length of 128, a batch size of 64 and a learning rate of 10^{-6}, the dropout probability is set to 0.1. We fine-tuned for 10 epochs, keeping the best model so far. We used the development set semEval2007(SE07) (Raganato et al., 2017) to fix the best parameters for our tests when fine-tuning. Concerning the final word representation, we average the final four layers of the first subword to get the representation of a word. That is, if a word is tokenized into a set of subwords, it's represented by the vector associated to the first subword.

For the English model, we use BERT with selection objective(Yap et al., 2020) and Gloss BERT (Huang et al., 2019) and (Du et al., 2019) that concatenate sentence embedding and gloss embedding for sentence representation. In addition, we tested with Roberta [1].

For the Arabic language, we use most available domain-specific pre-trained BERT models for modern standard Arabic (MSA): AraBERT (Antoun et al., 2020), Arabic-BERT (Safaya et al., 2020), CAMeL-BERT[2], and MARBERT[3] with arabGlossBERT (Al-Hajj and Jarrar, 2022). It's worthy mentioning that the multilingual mBERT (Libovickỳ et al., 2019) can also handle Arabic texts.

4.2 Datasets

For both languages, we carried out experiments on different corpora for data augmentation and for evaluation. Concerning data augmentation, our training is carried out on the SemCor corpus for the English language and its translated Arabic version for Arabic. Semcor is composed of 352 texts of the corpus Brown for a total of 226 040 annotations of meaning. This is the largest hand-annotated corpus available in English.

Regarding system evaluation, there are standard corpora for evaluation that are used or built up for evaluation campaigns.We use SE07 (Raganato

[1] https://huggingface.co/roberta-base

[2] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca

[3] https://huggingface.co/UBC-NLP/MARBERT

et al., 2017) for English and the corpus OntoNotes Release 5.0 (Weischedel et al., 2013) for Arabic language.

SemEval2007 (Navigli et al., 2007) dataset contains 5677 words for 2261 words annotated in WordNet by Raganato et al. (2017). OntoNotes Release 5.0 has three languages (English, Arabic and Chinese). It is a large corpus manually annotated in a legal sense containing several kinds of text (news, telephone conversations, weblogs, usenet newsgroups, broadcast, talk shows). The Arabic portion of OntoNotes Release 5.0 includes 300K words from the Arabic corpus An-Nahar Newswire. This corpus is available [4] and contains 212332 words for 12524 annotated senses from WordNet.

4.3 Results

As baseline, we used GlossBert and context BERT for English and ArabGlossBert for Arabic. Table 1 represents the results of our augmentation approach for the WSD task and a comparison for the English WSD with GlossBERT(Huang et al., 2019) and with ContextBERT(Du et al., 2019). Wei et al. (Wei and Zou, 2019) provided a list of easy and basic data augmentation techniques for text classification.

Concerning the English language, the proposed data augmentation technique outperforms the basic BERT model by 7% and achieves an accuracy score of 92.4% because DA adds diversity to the original dataset. We also note also that RoBERTa outperforms BERT for all the augmentation strategies.

Regarding the Arabic language, we obtain the best accuracy by following the context-tail augmentation strategy and by using the Arabic Bert embedding model. So, Arabic Bert with GPT2 based augmentation achieves a state of the art on Arabic WSD by an accuracy of 88.88% on Ontonote dataset (Salah et al., 2018b).

Words in sentences are known to have different parts and different levels of importance in the sentence. Thus, a particular word replacement or substitution has an effect different from that of replacing another word. A crucial issue arises concerning which word to select and replace it with another word or phrase. Table 1 answers this question and shows that, for any BERT-based model, the context-tail data augmentation outperforms the context-head and context-entire augmentation. This can be explained by the fact that replacing the tail of a sentence by a gloss is more sense preserving than the other two strategies.

[4] https://goo.gl/peHdKQ

	ModelDataAug	GPT-2: c-head	GPT-2:c-entire	GPT-2: c-tail	w/o
3*English Test: Sse-meval	BERT	90.20%	77.96%	91.16%	NA
	RoBERTa	91.63%	79.73%	**92.4%**	NA
	GlossBERT(BERT)B1	NA			80.0%
	ContextBERT(BERT)B2	NA			86.1%
8*Arabic Test: Ontonote	ArabGlossBERT(AraBERTv02)B3	NA			84%
8*	ArabGlossBERT(CAMeLBERT)B4	NA			82%
8*	ArabGlossBERT(QARiB)B5	NA			80%
	AraBERTv02	82.25%	65.38%	85.95%	NA
	Arabic BERT	85%	72.01%	**88.88%**	NA
	CAMeLBERT	84%	55%	86.22%	NA
	MARBERT	83.77%	57%	86%	NA
	mBERT	84.69%	59.64%	88.76%	NA

Table 1: Performances of data augmentation techniques for WSD. The final column refers to the basic model without any augmentation. NA stands for not applicable. w/o stands for without. Lines marked with a reference are experiments results from that reference. (Baselines B1 from the work implemented by (Huang et al., 2019), B2 from du2019using and B3, B4 and B5 from the ArabglossBERT (Al-Hajj and Jarrar, 2022)). The English models are tested on Semeval. The Arabic models are tested on Ontonote.

5 Conclusion

In this work, we presented a novel data augmentation framework based on the GPT2 model for word sense disambiguation. The generated training data is then combined with the gold dataset to train a BERT-based WSD classifier. The results of the experiment are very encouraging and show that the proposed contribution outperforms existing augmentation techniques. We also empirically proved that context-tail selection is the better context generation strategy.

As a future extension of this work, we aim to study the effects on other English or Arabic corpora as well as proposing different data augmentation techniques.

References

Moustafa Al-Hajj and Mustafa Jarrar. 2022. Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. *arXiv preprint arXiv:2205.09685*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Sabri Boughorbel, Fethi Jarray, Neethu Venugopal, and Haithum Elhadi. 2018. Alternating loss correction for preterm-birth prediction from ehr data with noisy labels. *arXiv preprint arXiv:1811.09782*.

Sabri Boughorbel, Fethi Jarray, Neethu Venugopal, Shabir Moosa, Haithum Elhadi, and Michel Makhlouf. 2019. Federated uncertainty-aware learning for distributed hospital ehr data. *arXiv preprint arXiv:1910.12191*.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using bert for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Harsh Kohli. 2021. Transfer learning and augmentation for word sense disambiguation. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 303–311. Springer.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. *arXiv preprint arXiv:2004.14769*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Guan-Ting Lin and Manuel Giambi. 2021. Context-gloss augmentation for improving word sense disambiguation. *arXiv preprint arXiv:2110.07174*.

William E Nagy. 1995. On the role of context in first- and second-language vocabulary learning. *Center for the Study of Reading Technical Report; no. 627.*

Roberto Navigli, Kenneth C Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, and Didier Schwab. 2018a. Un corpus en arabe annoté manuellement avec des sens wordnet. In *25e conférence sur le Traitement Automatique des Langues Naturelles*.

Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, and Didier Schwab. 2018b. Un corpus en arabe annoté manuellement avec des sens wordnet. In *25e conférence sur le Traitement Automatique des Langues Naturelles*.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. Ufsac: Unification of sense annotated corpora and tools. In *Language Resources and Evaluation Conference (LREC)*.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

R Weischedel, M Palmer, M Marcus, E Hovy, S Pradhan, L Ramshaw, N Xue, A Taylor, J Kaufman, M Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. web download. philadelphia: Linguistic data consortium, 2013.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting bert for word sense disambiguation with gloss selection objective and example sentences. *arXiv preprint arXiv:2009.11795*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Theorizing on Meaning-Making in Classroom Interaction Using "Critical Presuppositions-To-Theory" (C-PRETTY) Research

Janeson M. Miranda
De La Salle University Integrated School
Senior High School Division
janeson.miranda@dlsu.edu.ph

Rempson N. Gipaya
Marcelo H. Del Pilar National High School,
De La Salle University-Manila
rempson_gipaya@dlsu.edu.ph

Abstract

The paper proposes "Critical Presuppositions-To-Theory" (C-PRETTY) research as a new form of qualitative inquiry due to the realized challenges of grounded theory and the seemingly barren domain of theory building. The C-PRETTY serves as a special type of qualitative inquiry that focuses on *critical presuppositional framework* and *critical presuppositions* as the primary theoretical and methodological compass during the zigzagging data gathering and analysis. Guided by these, observations of two English classrooms, interviews with their teacher, and focus group discussions with the students were done to investigate classroom interaction with the belief that there is an exigency to create a contextualized theory to better depict the contextual realities of meaning-making as an integral part of classroom learning. The theory generated has been formalized as the *Multi-Layered Symbiotic Process of Meaning-Making,* which proposes that meaning-making starts at an Interactional Reference Point (IRP) and is carried through five mediums: locutionary, kinesic, affective-prosodic, cultural, and physical-spatial. These mediums interact with the intermediary layers as meaning travels through them. The findings of this study bring forth new ways of theorizing and conducting investigations, especially in the realm of educational linguistics.

Introduction

Despite the vibrancy and dynamism of qualitative researchers utilizing varying methods to strengthen the potency of investigation, only limited approaches or methods have been actualized to build theories in specific contexts of realities. In fact, Shah and Corley (2006) stated that although many journal editors have desired to cast more attention to theory development, fewer studies of such have been actually undertaken. Theories are regarded to be the "currency of scholarly research" (Corley & Gioia, 2011, p. 12). Therefore, new practical methods to theory building have to be explored and carried out to fortify the methodological efficiency and reliability of theory development, especially within the planes of qualitative inquiry.

Glaser and Strauss (1967), even decades ago, raised already a similar point, prompting them to initiate a notable way of theory building, which has been termed as 'grounded theory.' They expressed their dissatisfaction with the current existing theories that prevailed in sociological research and contended to move from data to theory so that new theories could potentially emanate. Their notion was eventually accepted and applied in several disciplines and later became a distinct type of qualitative study. It even permutated to several versions (MacDonald, 2001) depending on the researcher's ontological and epistemological perspectives (Mills et al., 2006).

However, several criticisms have been hurled against grounded theory and even Strauss and Glaser, who started out as colleagues have separated ways out of ontological and epistemological differences with regard to the method and data analysis.

One of the usual criticisms of grounded theory is the indefiniteness of the time when the researcher has finally achieved the theoretical or data saturation, an important feature of grounded theory (Aldiabat & Le Navenec, 2018). Theoretical saturation is achieved only when there have been no new patterns expected to emerge in dealing with the data (Charmaz, 2006; Glaser & Strauss, 1967) and that the rigors of the constant comparative analysis of data have been obviously exhausted (Glaser, 1992). This, according to Gasson (2004), is highly iterative and exhaustive. In the same vein, Hussein et al. (2014) pointed out grounded theory as having a potential for methodological errors, multiple approaches that may yield to confusion, and limited generalizability, and for its application of reviewing the literature without assumptions.

Thus, we endeavor to address some of these criticisms by proposing a new method or approach in theory building essentially called the "Critical Presuppositions-To-Theory" research or "C-PRETTY" approach to qualitative inquiry whose goal is analogous to grounded theory. However, the C-PRETTY research will attempt to attain the theoretical saturation by constantly refining the presuppositions after looking into the emerging patterns in data collection and analysis, which are conducted simultaneously. This is to efficaciously and consistently juxtapose the presuppositions with the data and continually refine them until such time that the presuppositions emerge to be (a) convincing theory(ies) after confirming and interpreting the clear connection or interplay between and among the categories, themes, or patterns inherently found in the data.

Theoretical saturation then is achieved in a definite phase when the critical presuppositions found to be truthful in their context and that the emerging themes or categories are at least lucidly auxiliary to these critical assumptions. Meaning, the presuppositions crystallize themselves in the light of the data and become an established theory in its own context. The C-PRETTY approach is a new way of qualitative inquiry, which targets to build a 'contextualized theory' out of critical presuppositions.

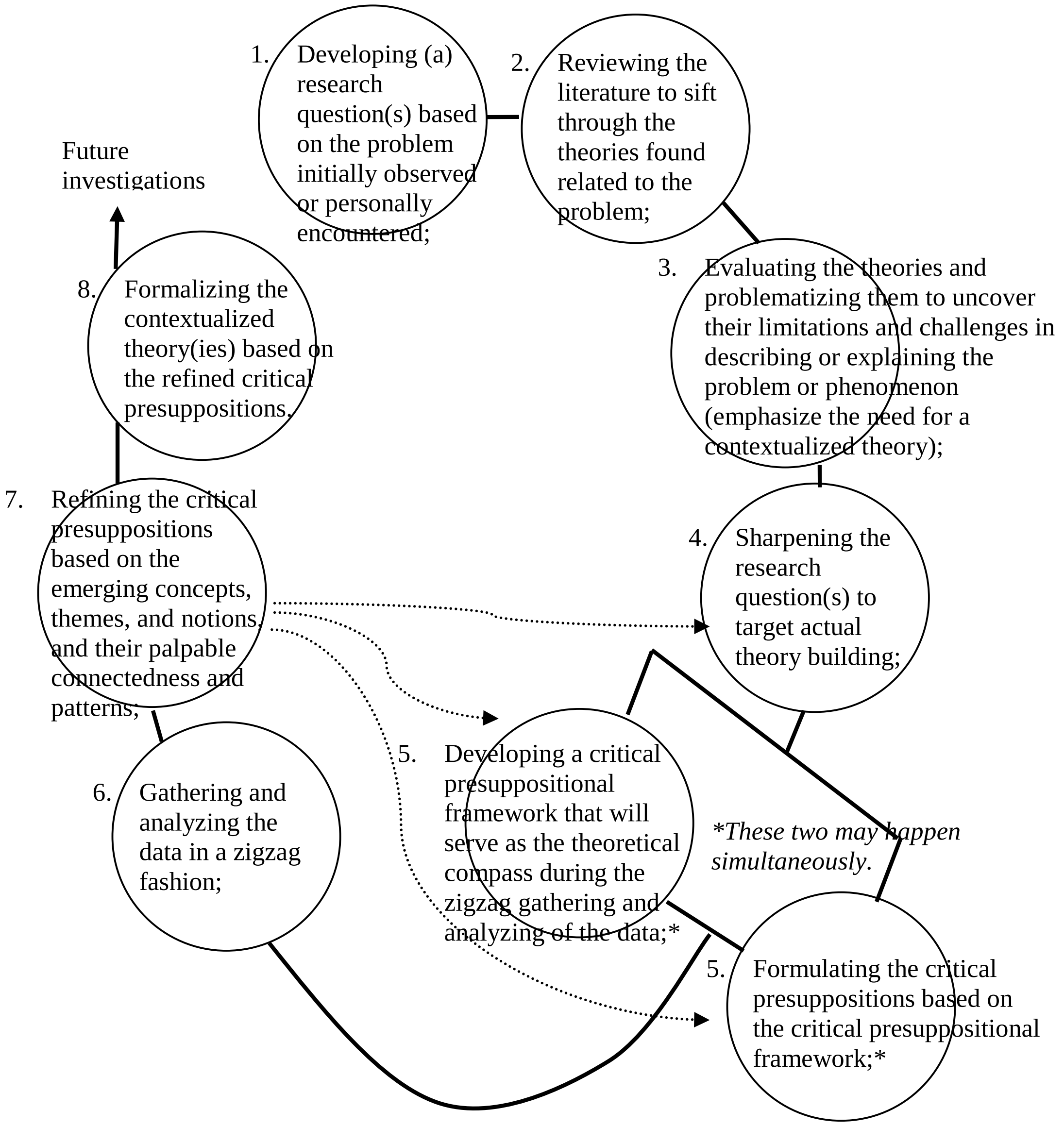
Contextualized theories, nevertheless, do not hold the notional universalities of everything in a specific orbit of human knowledge. As the term itself suggests—'contextualized'—a theory that will have been formed by the C-PRETTY research holds true in its own bounds or in its very context. Despite that, a contextualized theory does not restrict itself merely in its context and can transcend across disciplines and other areas of understanding. This can be done if a contextualized theory is utilized in another context in a future investigation and yields data and results that prove the theory's appropriateness and reliability. Should the theory transcend many fields of human psychological, social, philosophical, or even material knowledge only then will we behold the authentic capacity and utility of a theory; a contextualized theory then becomes a highly formalized theory with a sort of universal prowess to explicate broader realities or bigger human truths.

Crucial to this are critical presuppositions. Presuppositions are essential part of scientific investigations. Gay and Weaver (2011) argued that issues in research such as "definition, criteria, and purpose reflect an a priori commitment to certain presuppositional assumptions about what constitutes knowledge (epistemology), reality (metaphysics), the nature of being or existence (ontology), values (axiology), and other philosophical issues" (p. 24). In other words, theoretical presuppositions are but an inevitable part and parcel of any type of study that one wishes to undertake, be it quantitative or qualitative. All researchers are somewhat guided by a theoretical compass—consciously or unknowingly—as they navigate through the data or even at the onset of their scientific voyage.

Having explicated this, we maintain that the building of theory cannot come only from the data itself but from the interaction of the researcher with the data similar to what Charmaz (2000) posited in the light of the constructivist viewpoint on grounded theory. She stated that the data does not automatically offer a reality; rather the 'discovered' reality emerges from the interactive processes in its contextual factors (p. 524).

Hence, we further adhere to the belief that a researcher cannot come to the playing field blinded, unarmed, and incognizant of the surrounding circumstances of the issue they are investigating since the interaction with the data as what Charmaz (2002) suggested, can only be successful when the researcher is sufficiently informed. Having an informed mind before embarking on the actual data gathering in the C-PRETTY research means having an initial judicious evaluation of the problem or the subject in situ and attempting to find existing theories that will somehow capture a part of its reality, not its entirety. This is somewhat contradictory to what Glaser and Strauss (1967)

Figure 1
The C-PRETTY research process


1. Developing (a) research question(s) based on the problem initially observed or personally encountered;
2. Reviewing the literature to sift through the theories found related to the problem;
3. Evaluating the theories and problematizing them to uncover their limitations and challenges in describing or explaining the problem or phenomenon (emphasize the need for a contextualized theory);
4. Sharpening the research question(s) to target actual theory building;
5. Developing a critical presuppositional framework that will serve as the theoretical compass during the zigzag gathering and analyzing of the data;*
5. Formulating the critical presuppositions based on the critical presuppositional framework;*
*These two may happen simultaneously.
6. Gathering and analyzing the data in a zigzag fashion;
7. Refining the critical presuppositions based on the emerging concepts, themes, and notions, and their palpable connectedness and patterns;
8. Formalizing the contextualized theory(ies) based on the refined critical presuppositions.
Future investigations

and Corbin and Strauss (2008) advocated in grounded theory, which underscores the avoidance of pre-conceived notions of the situation prior to the data analysis to evade pre-cogitated theories. That is why we categorically distinguish the C-PRETTY research as a separate type of qualitative inquiry although its genesis can be traced back to the tenets of grounded theory, specifically to its methodological criticisms.

Clearly, what set the C-PRETTY research as a distinct lens of qualitative studies are the critical presuppositions and the manner how a theory is formed.

We define critical presuppositions as a paramount element of the C-PRETTY research achieved by looking at the problem in its actuality as an initial phase and checking it against the intensive perusal of related existing theories. The intensive perusal of existing theories in relation to the problem being examined will make the presuppositions *critical.* In other words, critical presuppositions must be a product of judicious understanding of the problem and possible applicable theories before indulging into the inductive-to-deductive interpretive/descriptive process when the researcher finally dives into the zigzag analysis of the actual data. However, the existing theories found to be related have to serve as the basis for problematization of such theories and thus prompt the impetus to question the theories' capacity and extent in describing or explaining the phenomenon. The problematization will then help the researcher point out the need for contextualized theories as they revise the initial research questions made. Problematization here, as argued, may question the institutionalized line of thinking and break the common notion of the majority (Alvesson & Sandberg, 2011, p. 32). Problematizing the existing theories and the other remaining steps for the C-PRETTY research process are indicated schematically in Figure 1.

Literature Review

Admittedly, one of the factors that birthed to C-PRETTY is our contention that most of the gargantuan theories in classroom interactions do not fully capture the realities of classroom communication, specifically the context of meaning-making. Although there are a number of investigations conducted that endeavored to study meaning-making in classroom contexts (e.g. Maarof & Yaacob, 2011; Axelsson & Slotte, 2017), such studies had only used the existing theories that will better explain classroom interactions. What such academic feats have been trying to realize is to continually apply the already existing and established theories in classroom interactional situations in specific actualities.

For example, Pardede (2017) investigated meaning-making of diversity of education students' experience during a 10-day intensive program employing a constructivist educational viewpoint put forward by notable theorist John Dewey. The researcher derived valuable understanding through the constructivist scaffold and presented a dialogic approach in meaning-making. However, what the study actualized was to find the classroom fit for the theory it used, but not really to create a new theoretical perspective of the issue being investigated.

Similarly, Maarof and Yaacob (2011) tapped in the theory of interactive reading and other existing reading theories in their investigation whereas Ganapathy et al. (2017) used theories on multimodal learning to scrutinize meaning-making in their own pedagogical sphere. None of them seemingly attempted to establish a certain theory that will fully capture the classroom interaction in their specific contexts, particularly focusing on meaning-making as an integral part of learning.

Meaning-making, as used in the literature for a long time, has yielded a number of definitions. One of which is Charanchi's (2016) definition; he defined meaning-making as an event which "involves mental activities and processes of constructing or deconstructing meaning of any linguistic aspect by the language learner" (p. 145). Such a definition is contextual and is specifically applicable in a language classroom. However, other definitions of meaning-making, which involve larger and complex interactional and pedagogical aspects can be found in the literature such as Barber's (2009) study which circumscribed meaning-making as a process of "connection, application, and synthesis" as part of "integrative learning" (p. 6).

As for the classroom interaction, insightful findings about it have been brought about by plethora of scholarly studies such as the investigations of Scott and Mortimer (2005), Kupferberg et al. (2009), Watanabe (2016), and Hamre et al. (2013) whose scope and rigor are perhaps the widest for they "test[ed] a developmental framework of teacher effectiveness in over 4,000 classrooms" in the light of "teaching through interactions" framework.

Clearly enough, the constructivist view of learning has become the primary theory that most academic researchers have taken into account for quite a long time and even heavily affected the methodological aspect of scientific inquiry.

For instance, the Flanders Interaction Analysis Categories (FIAC), which is also called as Flanders Interaction Analysis System (FIAS) has long gained the attention of many academicians and has been extensively used in the field to understand and evaluate the student and teacher interaction.

FIAS/FIAC has been used since its early recognition in many scientific investigations, which looked into the dimensions of classroom behavior (Medley & Hill, 1969), the verbal interaction between and among student teachers and educators (Smith, 1976), and even the interaction of physical education teachers with the students (Ritson et al., 1982). Attention to FIAS as a major framework for analysis has not waned and continuously gained the focus of recent scholarly investigations (e. g. See & Lim, 2006; Li et al., 2011; Amatari, 2015; Sharma, 2016). However, even before, Walker and Adelman (1975) already pointed out the limitations of FIAS in evaluating classroom interaction, especially of implicit “theory of instruction.” He argued that alternative methods of observation have to be developed to fully evaluate the meaning of interaction.

Apart from Flanders’ (1960) scholarly contribution, other methods of analysis have emerged that have served a great role in understanding the educative process. In the linguistic context, discourse analysis, conversation analysis, critical discourse analysis, and other pragmatically and socio-linguistically associated analyses have been applied to deepen and sharpen the scrutiny of classroom interactions, especially in line with meaning-making (see Maftoon & Shakouri, 2012; Sadeghi et al., 2012; Rogers et al., 2016).

Furthermore, another emerging perspective, which is strongly rooted from linguistics and has affected classroom praxis and ultimately the interaction that transpires therein, is the interactional sociolinguistics. Interactional sociolinguistics, as linked and associated to intercultural pragmatics, deals with “how language conveys meaning in interaction” (Tannen, 2005).

Tannen (2005) argued that researchers of interactional sociolinguistics tend to look into the intercultural interaction because meaning created through language out of this interaction appears lucidly on the surface. However, she noted that the role of linguistic processes in intercultural interaction as influenced by varied social factors shall also be taken into account.

Tannen’s (2005) position on interactional sociolinguistics is essential in our attempt to create critical presuppositions on classroom interactional framework in relation to meaning-making. This is so because if the meaning is constructed through language as a ‘dynamic’ and ‘emergent’ product of interaction among participants and not from a ‘single-handed linguistic production of individual speakers,’ we can actually regard classroom interaction as a dynamic and interconnected communicative event involving all the participants.

Apart from the interactional sociolinguistics, classroom interactions have long been investigated using several theoretical scaffolds such as Searle’s (1969) and Austin’s (1956) Speech Act Theory and Brown and Levinson’s Politeness Theory (1987). While it is true that such theories have helped explained and described classroom interactions in considerable ways, their limitations in capturing the actual classroom interaction especially in the context of meaning-making and learning cannot be altogether jettisoned.

For example, the locutionary, illocutionary, and perlocutionary concepts in Speech Act Theory, which concerns how meanings are formed in its expressive sense (locutionary), the motive or intention of the speaker (illocutionary), and the perception of and action done by the message receiver (perlocutionary), do not fully capture a classroom interactional situation. This is said so due to the fact that meaning-making can transpire as these three acts—locutionary, illocutionary, and perlocutionary—interact with one another and affect related meaning-making situation. For instance, when a language teacher says, “This is what you get for behaving the way you wanted to yesterday,” it may create multiple meanings in the minds of the students and a perlocutionary concept of one student may affect their classmates’ own perception. Assumingly, a student interprets it and whispers to John, “The teacher is angry,” the perlocutionary act or perception of the student then becomes another linguistic item for locutionary, illocutionary, and perlocutionary concept waiting for John’s personal interpretation and for other students who might have heard the whisper. John’s interpretation and of the ones who might have heard the whisper then will become another carrier of meaning expected to influence others’ perlocutionary acts if it is uttered in a speaking situation. This shows that meaning is negotiated as the participants and communicators interact and interchangeably take the role of the producer and receiver of messages.

Similarly, Politeness Theory which suggests that speakers have this ‘face’ which is defined as “a public self-image” and that ‘politeness’ is “showing awareness and consideration for another person’s face” (Yule, 2006, p. 119) seems to be problematic in many situations inside the classroom. The intended meaning of the teacher as constructed by their language is not consistently defined by the ‘face-threatening act’ or ‘face-saving act’ that they do in the actual teaching process. A teacher’s ‘face-threating act’ that may assert authority can just be deemed as nice and a seemingly similar thing with ‘face-saving act’ because of the consensual context and of the mutual cultural factors of both the students and the teacher. For example, a teacher’s utterance, "Here we go again! You're late! Sit down!" can be regarded clearly as a 'face-threatening act', but if the students just take it lightly and are cognizant that it is an inherent part of the teacher's personality, the utterance can be taken as a normal teacher's parlance and a non-threatening expression. The communicative meaning then is apparently arbitrarily cultural and contextual.

Such limitations of the aforementioned theories can be attributed to the fact that classroom interaction is a special type of communicative event different from a normal typical conversation since there are learning and teaching elements and features involved in its very fabric of dynamism. Further, since Speech Act Theory and Politeness Theory were not cogitated in the context of the classroom itself, the assumption that they will not fully and consistently capture the contextual realities of classroom interaction is but inevitable. A contextualized theory then is needed to better explicate classroom interactions. The stance of many to rely heavily on the existing theories not initially crystallized in the actual classroom set-up but will serve as a point of reference for pedagogical research should be exigently reevaluated.

Methodology

3.1 Research Design

As discussed, we employed the C-PRETTY research as the design of this study with the chief aim of generating a theory that will explain meaning-making that transpires in classroom interaction. Although the C-PRETTY is still in its infancy, its nature and aims are qualitative in essence and so will be classified as such.

Crucial to the design is the critical presuppositonal framework. Thus, having informed of the major notions and assumptions of classroom interactions in relation to meaning making, we constructed a critical presuppositional framework that served as a helpful scaffold in attaining initial critical presuppositions.

Even so, this framework evolved just as anticipated with the critical presuppositions during the actual data gathering and analysis. The framework chiefly served as our system of synthesizing and connecting conceptual and theoretical assumptions, which helped us in angling the investigation to have a more guided exploration and creation of a contextualized theory.

Table 1 succinctly indicates how we implemented the different stages of the C-PRETTY research along with some notes.

Table 1
Implementation of the C-PRETTY research as a method and approach

Stages		Implementation
1.	Developing (a) research question(s) based on the problem initially observed or personally encountered;	We formulated two initial research questions that focused on how meaning transpires in the classroom and how we can come up with a theory that will better explain meaning-making.
2.	Reviewing the literature to sift through the theories found related to the problem;	We tried to zero in on meaning-making as it happens inside the classroom. Keywords such as *interaction* were also found in the literature, which further aided our sharpening of research questions. Theories that were found were mostly linguistic in nature, such as the *interactional linguistics* among others.
3.	Evaluating the theories and problematizing them to uncover their limitations and challenges in describing or explaining the problem or phenomenon (emphasize the need for a contextualized theory);	We asked two major questions in problematizing each of the theories we had found: "Has this theory been formed in a classroom context? How does this theory fully capture classroom realities? Since we found that there is an apparent lack of theories that were generated in actual classroom contexts, we could now underscore the need to come up with a theory that is particularly focused on meaning-making and classroom interaction.

4.	Sharpening the research question(s) to target actual theory building;	After perusing the literature, we polished our research questions. We now asked, “How does meaning transpire through interaction in a language classroom?” and “What theory we can formulate that can best explicate this meaning-making in classroom interactions?”
5.	Developing a critical presuppositional framework that will serve as the theoretical compass during the zigzag gathering and analyzing of the data; and Formulating the critical presuppositions based on the critical presuppositional framework;	Informed by our personal experiences and guided by relevant ideas and insights derived from the literature, we devised a framework that initially explained meaning-making as it transpires in a classroom interaction. Then, we formulated critical presuppositions. The names or labels of the elements indicated in the framework were tentative and were expected to have changed during the actual data gathering and analysis. Developing a critical presuppositional framework and coming up with critical presuppositions can occur simultaneously or the theorizers can do one first before the other or vice-versa depending on how they deem to best realize them.
6.	Gathering and analyzing the data in a zigzag fashion;	We conducted first a classroom observation to initially investigate how meaning-making transpires in an actual classroom interaction and to determine the potential students who can participate in our focus group discussion (FGD). After this, interview data were sent to an intercoder for her independent analysis while we were also doing our own qualitative data analysis. Then, another classroom observation and batch of FGDs were done. Again, the data were sent to the same intercoder. After comparing our own data analysis with the intercoder’s analysis and determining the areas and aspects on which we have to focus more, we interviewed the teacher and proceeded with another classroom observation. In this manner, we were actually gathering data and immediately analyzing them—a zigzag fashion of collecting and analyzing data, which is essential in efficaciously refining the framework and critical presuppositions or even in possibly re-sharpening the research questions initially formulated. (For the detailed process, see the Procedure.)
7.	Refining the critical presuppositions based on the emerging concepts, themes, and notions, and their palpable connectedness and patterns;	In this stage, we continually refined the framework and the critical presuppositions based on the themes generated by us and our intercoder. We then gave a tentative name for the theory. The name was “Multi-layered process of meaning-making.”
8.	Formalizing the contextualized theory(ies) based on the refined critical presuppositions.	We finalized the name of the theory to be *“Multi-Layered Symbiotic Process of Meaning-Making*” to demonstrate the interplay of the elements found in the framework and the claims that have been put forward based on the critical presuppositions. We also consulted the two teachers involved in our study. We discussed with them the theory and employed a simple quantitative assessment of their confidence in the theory in explaining meaning-making in classroom interactions. This stage is called the Consensual Formalization of Theory. (For other details, see the Procedure.)

3.2 Participants

The participants in the study were two English language teachers and their classes; one taught Grade 8 class and the other Grade 6. At the time of data collection, the former class was composed of 28 students while the latter consisted of 24 students. Among the Grade 8 section, 15 of them participated in two batches of focus group discussions (FGDs). As to the Grade 6 class, 16 of them took part in two batches of FGDs. We purposively selected these learners based on their participation during the classroom observations conducted.

The students and teachers were from a small private school in Metro Manila in the Philippines.

3.3 Procedure

After securing necessary permission from the school authorities, explicit consent from the parents and the two teachers, and the students' assent, the first observation was done to Grade 6 English class. On the same day of the class observation, the first FGD was carried out with eight students. The selection of eight students was based on their participation, behavior, and performance during the class observation.

After the first FGD, the interview was translated and transcribed. The data then was given to the intercoder to look at the possible themes and categories to be coded. The inter-coder has been a teacher of English language for six years and was oriented first prior to the actual coding process.

The second observation was done to the same Grade 6 English class and after which, the second recorded FGD was undertaken with another set of eight students. The interview was translated and transcribed and sent again to the intercoder for her independent analysis.

After her initial data evaluation, a developed tool for determining the confidence of the intercoder in relation to the themes and categories in the data was accomplished. We formulated this tool to determine the "Percentage of Confidence (PoC)" of the intercoder and checked it against our own independent coding of the data. PoC requires the intercoder to rate each element or category in critical presuppositional framework based on some given ordinal criteria: '4' for strongly confident, '3' for confident, '2' for doubtful, and '1' for strongly doubtful. The answer for each category or element is then divided by the total highest possible score for all elements and is multiplied by 100 (e.g. 4/20 x 100). Finally, PoC is computed by getting the sum of all the elements in critical presuppositional framework. Thus, in this study, after checking the intercoder's independent data analysis against the elements in the critical presuppositional framework that we discussed with her, her level of confidence reached 85%, which is according to the developed tool can be regarded to be 'strongly confident.' The PoC range is as follows: 85-100= strongly confident; 75-84=confident; 50-74=slightly confident; 25-49=doubtful; 1-24=strongly doubtful.

After computing for the PoC, a one-on-one interview with the Grade 6 teacher was done. The 85% level of the intercoder's confidence prompted us also to continue the classroom observation by taking into consideration also the areas that the intercoder considered to be 'doubtful.' So, the third observation was done to Grade 8 English class, which followed by a recorded FGD with another set of eight students. Then, the transcribed and translated interview was sent to the intercoder for further evaluation while we also did our independent analysis. Finally, the fourth classroom observation and recorded FGD with seven students, and second one-on-one interview with the teacher were executed. The interview with the students and with the teacher was translated and transcribed, and was sent to the intercoder.

After the intercoder was done with the entire independent analysis and we also with our own data evaluation, we met with her and compared our codes and themes guided by the revised critical presuppositional framework. We discussed and we decided to collaboratively rate each element in the critical presuppositional framework using the PoC tool. The PoC now then yielded the Percentage of Consensual Confidence (PCC), the level of confidence emerged from our discussion with the intercoder. For this study, PCC reached 100%.

After reaching 100% of consensual confidence, we now discussed the critical presuppositional framework and the critical presuppositions with the two teachers who participated in the study. We also collaboratively rated each element in presuppositional framework and the critical presuppositions to get another PCC. In this process, a 100% of consensual confidence was reached anew. We propose this stage of C-PRETTy research as the Consensual Formalization of Theory; this is an essential stage as it actively involves the participants themselves in formalizing the theory per se.

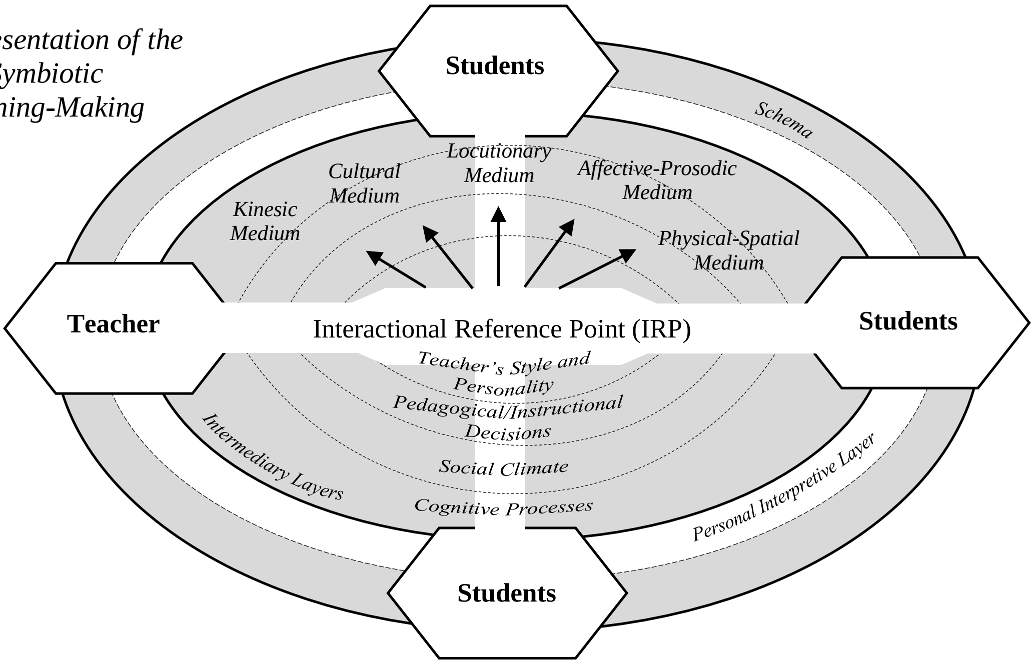
Results and Discussion

We have named the theory that was formalized in this C-PRETTY research as the *Multi-Layered Symbiotic Process of Meaning-Making* given the contextual symbioses of the different factors and elements and their connections with one another as explored and examined in the study. The framework of this theory is indicated in Figure 2.

The following are the theoretical claims of the *Multi-Layered Symbiotic Process of Meaning-Making*, which emanated from the critical presuppositions that were formulated, revised, and crystalized during the actual data gathering and analysis.

1. Every classroom teaching and learning communicative event has an Interactional Reference Point (IRP). The IRP refers to the starting plane of the conversation where either the teacher or the student performs an activity (linguistic or not or multi-modal) that serves as an impetus for continuous flow of interactional situation.
2. Meaning is carried through several mediums such as: (a) locutionary, (b) kinesic, (c) affective-prosodic, (d) cultural, and (e) physical-spatial. Locutionary medium refers to the actual words uttered by the speaker, which intrinsically carries meaning. These actual words are carried through the kinesic medium (gestures, facial expressions, body movements, etc.) and affective-prosodic medium which refers to the intonation, tone, juncture, and volume that carry the affective (feelings and emotions) of the speaker. Meaning then travels through the cultural medium from the speaker to the receiver of the message; cultural medium is the carrier of meaning on the layer of shared beliefs, practices, and understandings of the participants in the communicative situation. Finally, physical-spatial medium carries meaning through the present spatial context and proxemics of the participants and the entire physical dynamics of the interactional event.
3. These mediums that carry meaning then interact with the Intermediary Layers (ILs), which now form the interplay of complex factors of meaning-making in the teaching and learning process. These intermediary layers, which subtly or overtly influence meaning-making, can be outlined in a consequential and causal arrangement such as follows: teacher's style and personality → pedagogical/instructional decisions→ social climate→ cognitive processes→ meaning-making. We contend that the teacher's style and personality do determine their pedagogical/instructional decisions not the other way around. For instance, in the classes involved in our study, the personality of the teachers determines the type of decisions—instructional or social—which they make

Figure 2
Schematic representation of the Multi-Layered Symbiotic Process of Meaning-Making



in the classroom, such as correcting an answer through jokes, giving comments and suggestions, smiling, etc. It is not the pedagogical decisions that determine the personality or style of the teacher. Similarly, the pedagogical/instructional decisions coupled with the teacher's style and personality impact greatly the social climate of the classroom—whether it is accepting, positive, nurturing, or emotionally damaging. We likewise contend that the social climate profoundly impacts the student cognitive processes, which closely influence the personal meaning-making of the students.

4. Meaning therefore resides both in the mediums as it interacts with the intermediary layers and in the personal interpretive layer of the receiver of message. It travels through the interactions and does not hold fixed interpretion until it reaches the receiver's final interpretive layer. Hence, meaning is also a product of the interactive process of the message and the receiver's schema. Then, as shown in the framework, the meaning made in the personal interpretive layer may be used by the receiver as they continuously participate in the interaction in the Intermediary Layers (ILs) of the entire classroom communication.
5. ILs are crucial factors in meaning-making as they refer to the activities shared by both the teacher and the students—in other words, the interlocutors—as meaning travels through the mediums. They are pivotal since we critically presume that meaningful ILs create the intended meaning of learning.
6. Thus, meaning-making as an integral aspect of learning happens in a multi-layered fashion as the teacher and the students interact with one another creating intermediary layers that influence each student's personal interpretive layer.

Conclusions & Recommendations

After having endeavored to initiate the C-PRETTY research as a new qualitative way of generating a contextualized theory, we have concluded that it is possible to create theories in a specific context by addressing the limitations and criticisms of grounded theory without losing scientific rigor. With this, we have somehow democratized theory creation and made it more appealing even to neophyte or greenhorn theorists and researchers, such as undergraduate and master students since the C-PRETTY, if implemented systematically, can be comparably manageable, efficient, and practical.

As one implication of the theory generated to classroom instruction, teachers must also focus on the other mediums that carry meaning in the classroom especially the seemingly neglected elements such as kinesic, affective-prosodic, and even cultural mediums, alongside the intermediary layers. Equipping the teachers with good or even great teaching strategies is not enough as the question now is how these teaching strategies are implemented by the teacher and thus help build significant meaning-making in the classroom. There shall be a serious focus now, as implied by the theory, on the teachers' manner of using the mediums and the intermediary layers that carry meaning in the classroom since meaning is often associated with learning. The findings then in this study were relevant in potentiating our ways of theorizing especially in the field of educational linguistics.

To end, we put forward three recommendatory remarks. First, a continuous application of the C-PRETTY in many areas of learning or disciplines should be desired since we believe that the C-PRETTY is still on its refining stage; hence, suggestions and even criticisms should help its sharpening and development. Second, the C-PRETTY can be coupled with other quantitative research methodologies as one of the strategies in employing mixed-method type of research in order to stretch and determine the extent of its practicality, manageability, efficiency, and reliability both as a method and approach. And third, the theory of *Multi-Layered Symbiotic Process of Meaning-Making* may be employed in the same contexts such as the one within which the present study situates itself and even in other academic spheres in order to determine its usefulness or even weaknesses in capturing instructional and some general communicative truths and realities. This is due to the fact that the theory formed in this study, alongside the design and approach employed, is open to challenges and can still be refined by others.

Acknowledgments

We would like to give our profoundest thanks to Dr. Marianne Jennifer Gaerlan of De La Salle University for being one of the first who recognized the potential of the study during its proposal stage. The same gratitude goes to all the participants and individuals involved in the study—the teachers, students, our intercoder, and all the school authorities and parents who gave permission to carry out our investigation. Finally, we would like to thank the reviewers who rendered valuable feedback that helped improve the paper.

References

Aldiabat, K. M., & Le Navenec, C. (2018). Data saturation: The mysterious step in grounded theory method. *The Qualitative Report, 23*(1), 245-261. https://nsuworks.nova.edu/tqr/vol23/iss1/18

Alvesson, M. & Sandberg, J., 2011. Generating research questions through problematization. *Academy of Management Review*, *36*(2), 247-271.

Amatari, V.O. (2015). The instructional process: A review of Flanders' Interaction Analysis in a classroom setting. *International Journal of Secondary Education, 3*(5), 43 49. https://doi.org/10.11648/j.ijsedu.20150305.11

Austin, J. L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society*. Reprinted in Urmson, J.O. & Warnock G.J. (eds.). Philosophical Papers. Oxford University Press: Oxford 175-204.

Axelsson, M. & Slotte, A. (2017). Bridging academic and everyday language: Multilingual students' meaning-making in a lesson about Buddhism. *Journal of Immersion and Content-Based Language Education, 5*(2), 157-186. https://doi.org/10.1075/jicb.5.2.01axe

Barber, J. M. (2009). *Integration of learning: meaning making for undergraduates through connection, application, and synthesis* [Doctoral thesis, The University of Michigan]. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/62211/jpbarber_1.pdf;sequence=1

Brown, P. & Levinson, S. C. (1987) *Politeness: Some universals in language usage*. Cambridge University Press.

Charanchi, A. M. (2016). Meaning making strategies and challenges: Teaching English as a language in Nigerian educational institutions. *International Journal of Humanities and Management Sciences (IJHMS), 4*(2), 145-147.

Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 509-535). Sage.

Charmaz, K. (2002). Grounded theory analysis. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of Interview Research* (pp. 675–694). Sage.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. Sage Publications.

Corbin, J. & Strauss A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Sage

Corley, K. G., Gioia, D. A. (2011). Building theory about theory building: What constitutes a theoretical contribution?. *Academy of Management Review, 36*, 12-32

Flanders, N.A. (1960). *Interaction analysis in the classroom: A manual for observers*. University of Minnesota

Ganapathy, M., Saundravalli, A., & Seetharam, P. (2017). The effects of using multimodal approaches in meaning-making of 21st century literacy texts among ESL students in a private school in Malaysia. *Advances in Language & Literary Studies, 7*(2), 143-155. http://dx.doi.org/10.7575/aiac.alls.v.7n.2p.143

Gasson, S. (2004). Rigor in grounded theory research: An interpretive perspective on generating theory from qualitative field studies. In M. Whitman & A. Woszczynski (Eds.), *The Handbook of Information Systems Research* (pp, 79–102). IGI.

Gay, B. & Weaver, S. (2011). Theory building and paradigms: A primer on the nuances of theory construction. *American International Journal of Contemporary Research, 1*(2), 24-32. https://doi.org/10.1111/j.1469 - 5812.2007.00349.x

Glaser, B. G. & Strauss, A. (1967). *The discovery of grounded theory: Strategies for Qualitative Research*. Aldine.

Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs forcing.* Sociology Press

Hamre B. K., Pianta R.C., Downer, J.T., DeCoster J., Mashburn A. J., Jones, S. M., Brown J., Cappella, E., Atkins, A., Rivers, S.E., Brackett, M.A., &Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113*(4), 461-487.

Hussein, M. E., Hirst, S., Salyers, V., & Osuji, J. (2014). Using grounded theory as a method of inquiry: Advantages and disadvantages. *The Qualitative Report, 19*(27), 1-15. https://nsuworks.nova.edu/tqr/vol19/iss27/3

Kupferberg, I., Shimoni, S. & Vardi-Rath, E. (2009). Making sense of classroom interaction via a multiple-method design: social experiential and epistemological dimensions. *Linguagem em (Dis) curso—LemD., 9*(1), 81-106. https://doi.org/10.1590/S1518-76322009000100005

Li, L., Shouhui, Z. & Xinying, C. (2011, May 30-June 1). *Beyond research: Classroom interaction analysis techniques for classroom teachers* [Conference paper presentation]. 4th Redesigning Pedagogy International Conference, Singapore. https://repository.nie.edu.sg/bitstream/10497/6820/1/CRPP_2011_LiLi_a.pdf

Maftoon, P. & Shakouri, N. (2012). The concept of power in teacher talk: A critical discourse analysis. *World Applied Sciences Journal, 19* (8), 1208-1215, 2012 https://doi.org/10.5829/idosi.wasj.2012.19.08.1894.

Maarof, N., & Yaacob, M. (2011). Meaning-making in the first and second language: reading strategies of Malaysian students. *Procedia Social and Behavioral Sciences, 12,* 211– 223.

MacDonald, M. (2001). Finding a critical perspective in GT. In R. Schreiber & P. N. Stern (Eds.), *Using Grounded Theory in Nursing* (pp. 113-158). Springer.

Medley, D. & Hill. (1969). Dimensions of classroom behavior measured by two systems of interaction analysis. In ASCD Research Council (Eds.), *Research in Review*. http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_196905_medley.pdf

Mills, J., Bonner, A., & Francis, K. (2006). The development of constructivist grounded theory. *International Journal of Qualitative Methods, 5*(1), Article 3. http://www.ualberta.ca/~iiqm/backissues/5_1/pdf/mills.pdf

Pardede, S. (2017). *Understanding meaning-making of diversity: Education students' experience of a 10 day intensive programme* [Master's thesis, Department of Education University of Jyväskylä]. https://jyx.jyu.fi/dspace/bitstream/handle/123456789/55100/URN-NBN-fi-jyu-201708153488.pdf?sequence=5

Ritson, R. J., Smith, R.J. & Twa, H. I. (1982). Student and teacher interaction analysis: A comparison of activities, age groups and sex of the students in physical education. *Human Kinetics.* http://www.humankinetics.com/acucustom/sitename/Documents/DocumentItem/14916.pdf

Rogers R., Schaenen, I., Schott C., O'Brien K., Trigos-Carrillo L., Starkey, K., & Chasteen, C.C. (2016). Critical discourse analysis in education: A review of the literature, 2004 to 2012. *Review of Educational Research, 86*(4), 1192 – 1226. https://doi.org/10.3102/0034654316628993

Sadeghi, S., Ketabi, S., Tavakoli, M. & Sadeghi, M. (2012). Application of Critical Classroom Discourse Analysis (CCDA) in analyzing classroom interaction. *English Language Teaching,* (1). https://doi.org/10.5539/elt.v5n1p166

Scott P., Mortimer E. (2005). Meaning making in high school science classrooms: A framework for analysing meaning making interactions. In Boersma K., Goedhart M., de Jong O., Eijkelhof H. (Eds), *Research and the Quality of Science Education*. Springer, Dordrecht

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

See, K. H., & Lim, S. B. (2006). Effectiveness of interaction analysis feedback on the verbal behaviour of primary school mathematics teachers. *Malaysian Journal of Educators and Education, 21*, 115-128.

Shah, S.K. & Corley, K.G. (2006). Building better theory by bridging the quantitative-qualitative divide. *Journal of Management Studies, 43*(8), 1821-1835.

Sharma, S. (2016). A study of classroom interaction characteristics using Flander's classroom interaction analysis in a Maths class of rural and urban schools. *Scholarly Research Journal for Humanity Science and English Language, 3*(15) 3770-3776.

Smith, E.C. (1967). A latitudinal study of pre-service instruction in Flanders' Interaction Analysis Categories [Doctoral dissertation, Arizona State University]. https://files.eric.ed.gov/fulltext/ED120122.pdf

Tannen, D. (2005). Interactional Sociolinguistics as a resource for Intercultural Pragmatics. *Intercultural Pragmatics 2*(2) 205-208. DOI: 10.1515/iprg.2005.2.2.205

Walker, R. & Adelman, C. (1975). Interaction analysis in informal classrooms: a critical comment on the Flanders' system. *British Journal of Educational Psychology, 45*(1), 73-76. https://doi.org/10.1111/j.2044-8279.1975.tb02298.x

Watanabe, A. (2016). Engaging in an interactional routine in EFL classroom: the development of L2 interactional competence over time. *Novitas Royal Research on Youth and Language, 10*(1), 48-70.

Yule, G. (2002). *The study of language* (3rd ed.). Cambridge University Press.

Taste of Wine or “Taste” of a Person: (Synesthetic) Metaphors in Wine Reviews

Yin Zhong
Center for Language Education
The University of Science and Technology

lcyinzhong@ust.hk

Wing Sum Tse

Wong Shiu Chi Secondary School

edu121043@gmail.com

Abstract

Within the figurative repertoire of the genre of wine reviews, anthropomorphic metaphor is the most recurring pattern in many languages. However, few studies studied the metaphors in wine reviews in the Chinese language nor focused on the synesthetic metaphors as well as the synesthetic directionality in the wine discourse. This study built a small corpus of wine reviews and annotated conceptual metaphors, synesthetic metaphors, and synesthetic directionality in Chinese. With various metaphorical units concerning a person’s appearance and/or personality identified in our data, we confirmed the most frequent mapping of (TASTE OF) WINE IS A PERSON in Chinese wine reviews. In the meantime, VISION, which is considered a more abstract and less embodied sensory domain, however, provides the most vocabulary to describe the taste and smell of the wine. Although the finding seemingly violated the conventional mapping that more abstract concepts are comprehended via more concrete notions, we propose that the gustatory (and olfactory) sense is a “mutable” sensory domain in terms of its abstractness. Further, the evaluative similarity between the bodily experiences and the interactional communication may be the underlying cause of the reversibility of the TASTE and PERSONALITY as source domains in conceptual metaphor mappings.

Keywords: wine reviews, metaphors, Conceptual Metaphor Theory, synesthetic metaphors, linguistic synaesthesia

1 Introduction

Metaphors are ubiquitous in wine reviews, and they play a pivotal role in describing the wine tasting experience. Existing literature demonstrated that metaphor is a frequent and significant feature of the wine discourse (Caballero, 2007; Creed, 2016; Creed & McIlveen, 2018; Paradis & Eeg-Olofsson, 2013; Suárez-Toste, 2007). “Without metaphor, wine would be hard to discuss” (Caballero et al., 2019, p. 72). The prevalence of metaphor use in the genre of wine discourse may be partly because of the scarce taste and smell vocabulary in languages such as English (Levinson & Majid, 2014), although taste and smell are the primary two sensory faculties that take part in wine tasting.

To facilitate depicting wine tasting experiences, conceptual metaphorical frames are applied. For example, wine as a product will be portrayed as LIVING ORGANISMS (e.g., WINES ARE PLANTS; WINES ARE ANIMALS; WINES ARE PEOPLE), THREE-DIMENSIONAL ARTIFACTS (e.g., WINES ARE BUILDINGS, WINES ARE TEXTILES), and DYNAMIC ARTIFACTS (e.g., through manner-of-motion verbs like *run, ride, come across*) (Caballero et al., 2019; Creed, 2016). Another critical feature in winespeak[1] is that professionals and reviewers will use an array of sensory lexicons to evaluate wine attributes because tasting wine involves the activation of sensory perceptions via VISION, TASTE, SMELL, and TOUCH (or mouthfeel). VISION detects color and

[1] Winespeak refers to the specific terms or jargons that wine professionals use while discussing wine.

color depth, TASTE tells sweetness and acidity, SMELL distinguishes fruit intensity and oak presence, while TOUCH (or mouthfeel) evaluates the body, tannin, and carbonation of the wine (Old, 2014). These bodily sensations are closely intertwined and contribute to a holistic and integrated wine tasting experience. Linguistically speaking, meaning transfers across sensory domains are also considered a type of metaphor, namely, synesthetic metaphors. For example, in a phrase *sweet voice*, *sweet* is a concept originating in the taste sense while *voice* is a hearing concept—the auditory concept is thus depicted by the gustatory concept in this linguistic manifestation. Therefore, at least two perspectives can be offered in approaching the figurative device used in wine tasting discourse; one is through conceptual metaphors that associate wine with those more concrete or basic concepts, and the other is via synesthetic metaphors in which cross-mapping of sensory modalities is in pivotal interest.

Metaphors in wine discourse have received increasing attention in recent years and have been researched widely in languages such as English (Caballero et al., 2019; Creed, 2016; Paradis & Eeg-Olofsson, 2013), French (Negro, 2012), Italian (Ţenescu, 2014), Polish (Zawisławska & Falkowska, 2019), Spanish (Arroyo & Roberts, 2016), to name a few. Yet, little is known in regards to the (synesthetic) metaphors in the Chinese wine discourse, despite the fact that China is the sixth leading wine consumer worldwide, just after the United States, France, and Italy (Mercer, 2022). Wang et al. (2020) automatically extracted English-Chinese bilingual wine reviews and found asymmetric alignment between English and Chinese wine terms. Some frequently used words in wine reviews in English could not find their corresponding translation equivalents. For instance, palate, nose, and aromas could only be translated into 風格 *fēnggé* 'style'/口味 *kǒuwèi* 'taste,' 香氣 *xiāngqì* 'scent'/鼻腔 *bíqiāng* 'nasal cavity,' and 芬芳 *fēnfāng* 'fragrance' in Chinese, respectively. This leads us to question cross-linguistic and cross-cultural differences in lexical choices in wine reviews may also be reflected in metaphorical expressions.

This paper explores conceptual metaphors and synesthetic metaphors in wine reviews in Chinese. In particular, we address four questions in this research:

1) What are the frequently used conceptual metaphors in wine reviews?
2) What are the frequent source domains mapped to wine in wine reviews?
3) What are the frequently used synesthetic metaphors in wine reviews?
4) What is the synesthetic directionality in wine reviews?

We will delineate the theoretical framework in section 2; section 3 is on the methodology; results are presented in section 4, followed by discussions and conclusions in the last section.

2 Theoretical Framework

2.1 Conceptual Metaphor Theory

Metaphor is seen as a type of figurative device that describes one thing in terms of something else that is conceptually very different (Holyoak & Stamenković, 2018). One of the most influential accounts of metaphorical directionality is proposed by the Conceptual Metaphor Theory (CMT) (Lakoff & Johnson, 1980, 1999), in which conventional metaphorical expressions (e.g., *a warm person*) usually project a more concrete concept (e.g., TEMPERATURE) to a more abstract notion (e.g., SOCIAL RELATIONS). This cross-mapping typically involves a set of systematic correspondences underlying a conceptual metaphor (e.g., AFFECTION IS WARMTH) between the two conceptual domains (e.g., TEMPERATURE and SOCIAL RELATIONS). CMT advocates further posited that people would primarily draw upon their embodied experiences to comprehend metaphors (Gibbs, 2006; Gibbs et al., 2004) and reason about abstraction (e.g., Jamrozik et al., 2016).

In the context of wine discourse, past literature demonstrates that conceptualizations of the target domain of WINE were frequently found to arise from the ontological source domains of THREE-DIMENSIONAL ARTIFACTS, LIVING ORGANISMS, and MANUFACTURED ENTITIES. Among these source domains, the most pervasive metaphorical schema, irrespective of genre and wine community, regards (TASTE OF) WINE as A PERSON (e.g., Caballero et al., 2019; Creed & McIlveen, 2018; Suárez-Toste, 2007). Suárez-Toste (2007) described that wine's *personality* is evaluated by means of adjectives prototypically used in the qualification of human beings (e.g., *brooding*, *friendly*, *sexy*, *voluptuous, boisterous*, *assertive*, *sensitive*, *demure*, *shy*, or *expressive*). Creed (2016, p. 152) also found that there was a strong connection between human personality traits involving behavior and characteristics (e.g., *brooding, character, clever, generous, gentle, honest*, and *mellow*) and physical actions (e.g., *clamoring, demanding, promising, shows*, and *sings*).

In this sense, the source domain of PERSON or PERSONALITY in conceptualizing the target domain of TASTE and/or SMELL of the wine seems to violate the conventional mappings in conceptual metaphors, i.e., using more concrete ideas to facilitate understanding of more abstract concepts, given that flavor and odor are more concrete and more "embodied" than getting to know a person or evaluating a person's traits. Can we really understand a *sexy* or a *clever* taste? We will come back to this issue in the discussion section.

2.2 Synesthetic Metaphors and their Directionality

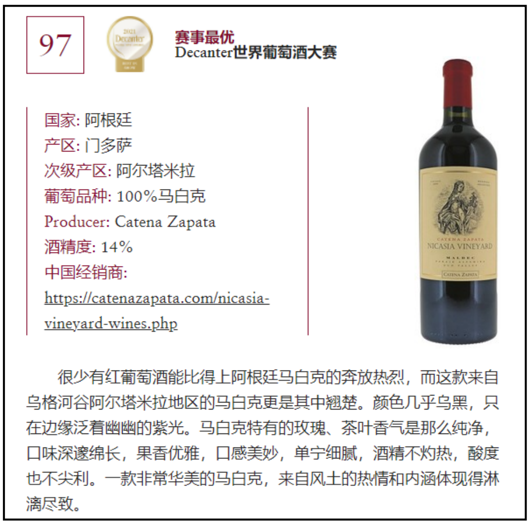
Synesthetic metaphors specifically refer to cross-sensory metaphors that involve two sensory domains, i.e., VISION, HEARING, TASTE, SMELL, and TOUCH. Similar to the typical transfer pattern in the CMT, i.e., from a more concrete concept to a more abstract concept, synesthetic directionality from this metaphoric point of view likewise follows a particular directional pattern, from a "lower," or a "more embodied" sense (e.g., TOUCH) to a "higher," or a "less embodied" sense (e.g., HEARING) (see Strik Lievers et al., 2021 for a summary). The differentiation of "lower" and "higher" senses primarily lies in their degrees of embodiment, more specifically, the involvement and closeness of bodily contact (Shen, 1997; Shen & Aisenman, 2008) as well as with reference to subjective and/or objective information (Popova, 2005). For instance, tactile (and gustatory) senses appearing at initial mapping points to other senses (visual and auditory senses) can be explained by their substantial involvement of bodily contact and references to subjective feelings rather than objective information perceived by visual and auditory senses. Linguistic synesthesia is thus conventionally approached as a type of conceptual metaphor (Shen, 1997; Strik Lievers, 2017; Yu, 2003; Zhao et al., 2019a) and lexical items concerning meaning transfers are termed "synesthetic metaphors."

Suárez-Toste (2017) presented a case in point regarding the acidity in wine. He found that the directionality of synesthetic metaphors in discussing the acidity violated Ullmann's (1957) or Williams' (1976) synesthetic hierarchy in which TASTE is always found transferring to VISION. He identified a wealth of visual terms, especially related to light, such as *bright, beam, shine, shimmering, vivid, clarity, streak, laser, flashlight,* etc., were employed to describe the acidity in wine. This finding surprisingly echoes the mapping directionality in conceptual metaphors, as mentioned in the above section. Does it suggest that TASTE and SMELL are more abstract and less "embodied" than VISION? More discussions on this point can be found in the discussion after we present the findings of this study.

3 Methods

3.1 Data

The data used in this paper were collected from Decanter China, www.decanterchina.com. The data source consists of 50 wine reviews for wines that have been awarded the 2021 Decanter World Wine Award (DWWA). A sample of the review is demonstrated in Figure 1.



97

赛事最优
Decanter世界葡萄酒大赛

国家: 阿根廷
产区: 门多萨
次级产区: 阿尔塔米拉
葡萄品种: 100%马白克
Producer: Catena Zapata
酒精度: 14%
中国经销商:
https://catenazapata.com/nicasia-vineyard-wines.php

很少有红葡萄酒能比得上阿根廷马白克的奔放热烈，而这款来自乌格河谷阿尔塔米拉地区的马白克更是其中翘楚。颜色几乎乌黑，只在边缘泛着幽幽的紫光。马白克特有的玫瑰、茶叶香气是那么纯净，口味深邃绵长，果香优雅，口感美妙，单宁细腻，酒精不灼热，酸度也不尖利。一款非常华美的马白克，来自风土的热情和内涵体现得淋漓尽致。

Figure 1. *A sample of the wine review in Chinese*

As shown in Figure 1, a wine review usually includes a technical introduction listing the wine name, vintage year, country/region/sub-region of origin, type of grapes, producer, and concentration of alcohol. The descriptions underneath typically start with a general introduction and evaluation of the wine, followed by important attributes centered on the wine, including color, aroma, flavor, body, and tannin, and ends with the wine reviewer's overall appraisal.

Since we are only interested in the metaphor used in the reviews, we thus only extracted the descriptions for each review. The facts and information about the wine were discarded in this paper. After cleaning the data, we have compiled a small wine review corpus with 9,477 Chinese characters for the 50 wines.

3.2 Procedure

Two coders with trained linguistic background were involved in the following procedures of identifying metaphorical units, source and target domains coding, as well as sensory lexicon categorizations. All the questionable cases were discussed and resolved between the coders.

3.2.1 Metaphor Identification

We mainly resorted to a bottom-up approach and annotated the data manually, i.e., without previous automatic pre-detection of keywords (Stefanowitsch & Gries, 2006). This is because we would like to observe the data in a more exploratory way.

First, we adopted the Metaphor Identification Procedure VU University Amsterdam (MIPVU; Pragglejaz, 2007; Steen et al., 2010) to identify conceptual metaphors. More specifically:

a. To read the entire text and to get a general understanding of the meaning;
b. Determine the lexical unit;
c. Establish the contextual meaning of each lexical unit
d. Determine if the word has a more basic meaning (more concrete, more bodily-related, more precise, and more historically older) in other contexts than the one in the given context;
e. If the lexical unit has a more basic meaning in other contexts than the given context, to decide if the contextual meaning contrasts with the basic meaning but can be understood in comparison with it;
f. If yes, mark the lexical unit as a metaphor-related word.

(Pragglejaz, 2007, p. 3; Steen et al., 2010, pp. 5-6)

Three lexical tools were used to determine the word meanings in Chinese, *Chinese WordNet 2.0* (CWN, Huang et al., 2010)[2] and two reference dictionaries, *Handian*,[3] and *Xiandai Hanyu Cidian* (The Contemporary Chinese Dictionary, Dictionary Editing Office, 2016). The three lexical tools will complement each other in determining the contextual and more basic meaning of the word we concern with. We generally followed the identification criteria for the Chinese data in Lu and Wang (2017) and Tay (2015). First, we only considered the basic meaning of the entire compound rather than that for the single character. For example, we treated the compound word 細膩 *xìnì* as one lexical unit with the meaning of '*fine and smooth*' instead of analyzing the basic meaning of the two characters 細 *xì* 'thin' and 膩 *nì* 'greasy; excessively (flavored).' Secondly, we included similes, idioms, colloquialisms, and proverbs that involve metaphorical meanings. Lastly, we took metonymy into account.

3.2.2 Source-Target Domains Coding

Three steps in the verification of source domains were adopted (Ahrens & Jiang, 2020; Zeng et al., 2021):

a. To propose a potential source domain based on educated and native speakers' judgment, accompanied by the co-text and context of the metaphorical word appears;
b. To verify the source domain proposed in the first step by checking if the categories and meanings of the metaphorical words provided in CWN, two dictionaries mentioned above, and two ontological knowledge networks (i.e., *E-HowNet* (Ma & Shih, 2018) and *SUMO* (Suggested Upper Merged Ontology) (Niles & Pease, 2001)) relate to the proposed source domain;
c. If no evidence can be found in (b), the collocation searches of the keywords in the *Sketch Engine* (Kilgarriff et al., 2014) will be checked to examine if there are any

[2] CWN is a platform provides an ontological network of semantic meanings of a word coupled with their semantic relations, including hypernyms, hyponyms, synonyms, among others. Accessed at http://lope.linguistics.ntu.edu.tw/cwn2/.

[3] An online Chinese dictionary; accessed at https://www.zdic.net/.

frequent collocations of the keywords related to the proposed source domains.[4]

As for the target domains, we mainly read through the whole sentence and analyze the target issues that the metaphorical words relate to.

3.2.3 *Synesthetic Metaphors Coding*

The boundary of sensory domains can be fuzzy. For example, the sensory adjective 清爽 *qīngshuǎng* ‘refreshing’ can be used to denote auditory, gustatory, olfactory, and tactile feelings, although its original meaning is more pertinent to the visual sense. The classification of the sensory vocabulary mainly follows the method of categorizing sensory words in Zhong and Huang (2020) and Zhong et al. (2022a):

a. The sensory domain that the etymology of the word is pertinent to (cf. Zhao et al., 2019b);
b. The dominant sensory domain that the word belongs to (cf. Chen et al., 2019; Zhong et al., 2022b);
c. Other words not listed in a or b will be traced in Shuowenjiezi (Xu, 1963) for their original connotations or their frequent usages in a general corpus data in the Sketch Engine, e.g., Chinese Web 2017 (zhTenTen11).

4 Results

After reviewing all the 50 wine reviews written of the awarded wine in DWWA 2021 and applying the above coding methods, we have identified a total of 345 metaphorical instances that contained metaphorical keywords. 151 of them (which takes 43.8%) are synesthetic metaphors, while the remaining 194 expressions (accounts for 56.2%) are conceptual metaphors. We will explicate the findings in the following two sections.

4.1 Conceptual Metaphors in Wine Reviews

On metaphors in general (excluding synesthetic metaphors), the most frequent metaphorical words include 優雅 *yōuyǎ* ‘elegant’ (4.06%), 充沛 *chōngpèi* ‘abundant’ (3.19%), and 內斂 *nèiliǎn* ‘introverted’ (2.61%). Table 1 demonstrates the top 10 metaphorical keywords identified in wine reviews.

Metaphorical Keywords	**Frequency (Percentage)**
優雅 *yōuyǎ* ‘elegant’	14 (4.06%)
充沛 *chōngpèi* ‘abundant’	11 (3.19%)
內斂 *nèiliǎn* ‘introverted’	9 (2.61%)
活潑 *huópō* ‘lively’	8 (2.32%)
迷人 *mírén* ‘charming’	8 (2.32%)
年輕 *niánqīng* ‘young’	7 (2.03%)
個性 *gèxìng* ‘personality’	6 (1.74%)
層次 *céngcì* ‘layer’	6 (1.74%)
和諧 *héxié* ‘harmonious’	5 (1.45%)
輕盈 *qīngyíng* ‘light’	5 (1.45%)

Table 1. *Top 10 most frequent metaphorical keywords in wine reviews*

Most of the source domains lie in PERSON (76.8%), BUILDING (9.7%), FORCE (4.1%), WATER (4.1%), ATMOSPHERE (2.5%), and STORM (1%). While the target domains mainly involve the TASTE of the wine (57.2%), the wine *per se* (26.2%), the SMELL of the wine (15.4%), and the COLOR of the wine (1%). We can therefore establish the frequent conceptual metaphorical mapping in wine reviews is (TASTE OF) WINE IS A PERSON.

4.2 Synesthetic Metaphors in Wine Reviews

Concerning synesthetic metaphors in wine reviews, most of the source domains lie in VISION (60.9%), TOUCH (25.1%), and HEARING (9.9%). According to Table 2, the most frequent synesthetic metaphors include 柔和 *róuhé* ‘soft’ (5.22%)，清新 *qīngxīn* ‘fresh’ (4.06%), and 純淨 *chúnjìng* ‘pure’ (4.06%).

Synesthetic Metaphors	**Frequency (Percentage)**
柔和 *róuhé* ‘soft’	18 (5.22%)
清新 *qīngxīn* ‘fresh’	14 (4.06%)
純淨 *chúnjìng* ‘pure’	14 (4.06%)
細膩 *xìnì* ‘fine and smooth’	14 (4.06%)
深沉 *shēnchén* ‘deep’	15 (4.35%)
清爽 *qīngshuǎng* ‘refreshing’	10 (2.80%)
柔滑 *róuhuá* ‘silky’	5 (1.45%)
深邃 *shēnsuì* ‘deep’	5 (1.45%)
豐美 *fēngměi* ‘plump’	5 (1.45%)

[4] The Sketch Engine is accessed at https://the.sketchengine.co.uk/auth/corpora/.

乾淨 *gānjìng* ‘clean’	3 (0.870%)

Table 2. *Top 10 most frequent synesthetic metaphors in wine reviews*

On top of the synesthetic metaphors, we further summarized the synesthetic directionality in wine reviews, as presented in Figure 2. The synesthetic metaphors follow a particular direction, in which unidirectional is most seen, while the bidirectional transfer is only found between TASTE and SMELL. Taking visual items to describe the TASTE is the most often among the mapping to the taste of the wine, which takes 57.4%. It is also observable that TOUCH is also often used in describing TASTE, as it takes 32.4%. HEARING is the rarest sense used to describe the TASTE (7.4%). When the SMELL of the wine is the target, VISION is likewise the most frequent sensory modality mapped to the SMELL (72%), followed by TOUCH to SMELL (12%) and HEARING to SMELL (4%). Last but not least, only the auditory terms are found to modify the VISION, more specifically, the color of the wine.

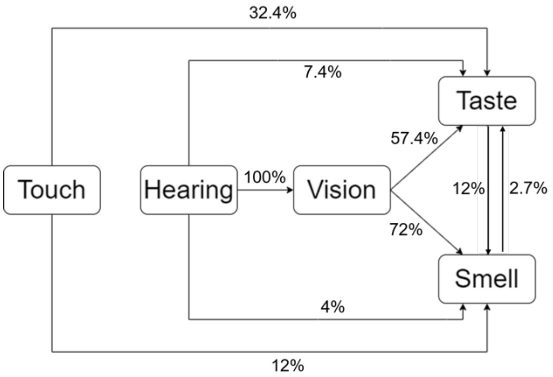


Figure 2. *Synesthetic directionality in wine reviews*

5 Discussion

The conceptual metaphors and synesthetic metaphors in wine reviews in Chinese exhibit an analogous pattern to those found in other languages. Personification, or anthropomorphic metaphors, predominate the wine discourse; and VISION, deemed a less “embodied” sense, is the most frequent sensory domain that transfers its meaning to TASTE as well as SMELL in wine reviews. The reversibility of the source and target domains in both the conceptual metaphorical mappings and synesthetic metaphors left us with the question of the “abstractness” of the gustatory and the olfactory sensory domains—can we really understand what is the “taste” or “smell” of a *sexy* or a *clever* person? Does it suggest that TASTE and SMELL are more abstract and less “embodied” than VISION?

To kill two birds with one stone, we would like to corroborate the hypothesis that the gustatory (and olfactory) category, although bodily experienced, is conceptually abstract (Zhong & Huang, 2018, 2020; Zhong et al., 2022a). Based on the analysis of online food reviews on desserts, Zhong and Huang (2018, 2020) suggested that the “mouthfeel” of desserts is also conceptualized as an individual’s personality since the adjectives stemming from impressions of personalities gained through social interactional contact were found modifying the taste as well as the “mouthfeel” of desserts in Chinese, such as 調皮 *tiáopí* ‘naughty,’ 浮誇 *fúkuā* ‘superficial,’ and 硬朗 *yìnglǎng* ‘robust.’ In Zhong et al.’s (2022a) examination of the adjectives collocated with the gustatory/olfactory word 味道 *wèidào* ‘taste; smell’ using the corpus data, they found that visual-related adjectives, especially those related to simplicity, purity, and elegance (e.g., 純 *chún* ‘pure,’ 淡雅 *dànyǎ* ‘simple and elegant,’ and 精緻 *jīngzhì* ‘delicate’), can be used to modify 味道 *wèidào* ‘taste; smell’ in Chinese. They further hypothesized that when the focus is on the quality of sensation, i.e., the desirability and pleasantness of taste, it is reasonable to use more conceptual terms to modify the target items.

The high consistency in conceptualizing TASTE and SMELL through personification shows that this phenomenon is not exclusively unique in the genre of wine reviews. It is neither because wine is anthropomorphic nor bears humanlike characteristics by nature (Creed & McIlveen, 2018). We hypothesize that the mutability of the TASTE and SMELL on the concrete-abstract continuum might be the underlying reason that drives the arising of the PERSON or PERSONALITY to be the most common source domain in the wine discourse. Further, personality-related lexical items carry evaluative appraisal through interactional contact with the people, which is tantamount to the evaluative function of gustatory and/or olfactory terms used to describe personal experiences with the things (Winter, 2016). This also helps explain why TASTE and SMELL will drift towards a more abstract end when the appraisal of the quality is the primary concern. Overall speaking, we propose that the “preexisting similarity” between the two experiences (Kövecses, 2002) grounds the projection of the “taste” of a person to the taste of the wine.

6 Conclusion

Describing the taste and smell of the wine often requires costly and creative endeavors from wine professionals because people generally have difficulties in naming flavors and odors, especially in Western cultures (Croijmans & Majid, 2015; Croijmans et al., 2021; Levinson & Majid, 2014). However, studies also showed that wine experts did not demonstrate much difference in using linguistic strategies to communicate smells and flavors (Croijmans & Majid, 2016), although they will employ more metaphorical descriptions to describe wine (Paradis & Eeg-Olofsson, 2013). This study thus looks at how conceptual metaphors and synesthetic metaphors are used in wine reviews in Chinese. Our findings echo the past literature on the metaphorical expressions in wine discourse in other languages. First, a variety of metaphorical units concerning a person’s appearance or a person’s personality are identified, which further gives rise to the most frequent mapping of (TASTE OF) WINE IS A PERSON in the winespeak. Secondly, VISION, which is considered a more abstract and less embodied sensory domain in conventional linguistic synesthesia, provides the most vocabulary to describe the taste and smell of the wine.

The findings in the genre of wine reviews somewhat share a remarkable similarity with taste/smell descriptors in a general sense. We hypothesize that the gustatory (and olfactory) sense is a “mutable” sensory domain in terms of its abstractness. Especially, the conceptual abstractness of the TASTE (and SMELL) is activated when the quality of bodily sensation is the focus. Further, the evaluative similarity between the bodily experiences and the interactional communication may be the underlying cause of the reversibility of the TASTE and PERSONALITY as source domains in conceptual metaphor mappings.

Due to space limitations, we did not elaborate on other source domains and/or metaphorical mappings, such as (TASTE OF) WINE IS A BUILDING. The small size of the corpus may also limit our findings of other possible yet novel or creative metaphor usages in Chinese. Future studies may use experimental methods to test the acceptability of different aspects and features related to a person in discussing taste and smell—for example, if a person’s psychological attribute is more acceptable than their physical features in winespeak. With the consistent patterns identified in metaphor use, it is also worthwhile to resort to computational models to predict the figurative language in wine discourse.

References

Ahrens, K., & Jiang, M. (2020). Source domain verification using corpus-based tools. *Metaphor and Symbol*, *35*(1), 43-55. https://doi.org/10.1080/10926488.2020.1712783

Arroyo, B. L., & Roberts, R. P. (2016). Differences in wine tasting notes in English and Spanish. *Babel (Frankfurt)*, *62*(3), 370-401. https://doi.org/10.1075/babel.62.3.02lop

Caballero, R. (2007). Manner-of-motion verbs in wine description. *Journal of Pragmatics*, *39*(12), 2095-2114. https://doi.org/10.1016/j.pragma.2007.07.005

Caballero, R., Suárez-Toste, E., & Paradis, C. (2019). *Representing wine: sensory perceptions, communication and cultures*. John Benjamins Publishing Company.

Chen, I.-H., Zhao, Q., Long, Y., Lu, Q., & Huang, C.-R. (2019). Mandarin Chinese modality exclusivity norms. *PLoS ONE*, *14*(2), e0211336. https://doi.org/10.1371/journal.pone.0211336

Creed, A. (2016). *Wine communication in a global market: A study of metaphor through the genre of Australian wine reviews.* [Doctoral thesis, University of Southern Queensland].

Creed, A., & McIlveen, P. (2018). Uncorking the potential of wine language for young wine tourists. In M. Sigala & R. N. S. Robinson (Eds.), *Management and marketing of wine tourism businesses: Theory, practice and cases* (pp. 25-41). Springer International Publishing. https://doi.org/10.1007/978-3-319-75462-8_2

Croijmans, I., & Majid, A. (2015). *Odor naming is difficult, even for wine and coffee experts*.

Croijmans, I., & Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *PLoS ONE*, *11*(6), e0155845-e0155845. https://doi.org/10.1371/journal.pone.0155845

Croijmans, I. M., Arshamian, A., Speed, L. J., & Majid, A. (2021). Wine experts' recognition of wine odors is not verbally mediated. *Journal of Experimental Psychology. General*, *150*(3), 545-559. https://doi.org/10.1037/xge0000949

Dictionary Editing Office, I. o. L., Chinese Academy of Social Sciences. (2016). *Xiandai Hanyu Cidian [The Contemporary Chinese Dictionary]* (C. A. o. S. S. Dictionary Editorial Office of Language Institute, Ed. 7th ed.). The Commercial Press.

Gibbs, R. W. (2006). Metaphor interpretation as embodied simulation. *Mind & Language*, *21*(3), 434-458. https://doi.org/10.1111/j.1468-0017.2006.00285.x

Gibbs, R. W., Costa Lima, P. L., & Francozo, E. (2004). Metaphor is grounded in embodied experience. *Journal of Pragmatics*, *36*(7), 1189-1210. https://doi.org/10.1016/j.pragma.2003.10.009

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641-671. https://doi.org/10.1037/bul0000145

Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., & Huang, S.-W. (2010). Chinese Wordnet: design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, *24*(2), 14-23.

Jamrozik, A., Jamrozik, A., McQuire, M., McQuire, M., Cardillo, E. R., Cardillo, E. R., Chatterjee, A., & Chatterjee, A. (2016). Metaphor: Bridging embodiment to abstraction. *Psychonomic Bulletin & Review*, *23*(4), 1080-1089. https://doi.org/10.3758/s13423-015-0861-0

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Kövecses, Z. (2002). *Metaphor: a practical introduction*. Oxford University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language*, *29*(4), 407-427. https://doi.org/10.1111/mila.12057

Lu, X., & Wang, B. P.-Y. (2017). Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, *51*(3), 663-694. https://doi.org/10.1007/s10579-017-9392-9

Ma, W.-Y., & Shih, Y.-Y. (2018). Extended HowNet 2.0 - an entity-relation common-sense representation model. 11th International Conference on Language Resources and Evaluation (LREC-11), Miyazaki, Japan.

Mercer, C. (2022). *Which countries drink the most wine? Ask Decanter*. Retrieved 17 May 2022 from https://www.decanter.com/wine-news/which-countries-drink-the-most-wine-ask-decanter-456922/

Negro, I. (2012). Wine discourse in the French language. *Revista Espanola de Linguistica Aplicada*, *11*(11), 1-12.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *International Conference on Formal Ontology in Information Systems*, 2-9.

Old, M. (2014). *Wine: a tasting course*. DK Publishing.

Paradis, C., & Eeg-Olofsson, M. (2013). Describing sensory experience: the genre of wine reviews. *Metaphor and Symbol*, *28*(1), 22-40. https://doi.org/10.1080/10926488.2013.742838

Popova, Y. (2005). Image schemas and verbal synaesthesia. In B. Hampe & J. E. Grady (Eds.), *From perception to meaning: image schemas in cognitive linguistics* (pp. 395-420). Mouton de Gruyter. https://doi.org/10.1515/9783110197532.5.395

Pragglejaz, G. (2007). MIP: a method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*(1), 1-39. https://doi.org/10.1207/s15327868ms2201_1

Shen, Y. (1997). Cognitive constraints on poetic figures. *Cognitive Linguistics*, *8*(1), 33-72. https://doi.org/10.1515/cogl.1997.8.1.33

Shen, Y., & Aisenman, R. (2008). 'Heard melodies are sweet, but those unheard are sweeter': synaesthetic metaphors and cognition. *Language and Literature*, *17*(2), 107-121. https://doi.org/10.1177/0963947007088222

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A* *method for linguistic metaphor identification: from MIP to MIPVU*. John Benjamins Pub. Co.

Stefanowitsch, A., & Gries, S. T. (2006). *Corpus-based approaches to metaphor and metonymy*. M. de Gruyter.

Strik Lievers, F. (2017). Figures and the senses: Towards a definition of synaesthesia. *Review of Cognitive Linguistics*, *15*(1), 83-101. https://doi.org/10.1075/rcl.15.1.04str

Strik Lievers, F., Huang, C.-R., & Xiong, J. J. (2021). Linguistic synaesthesia. In X. Wen & J. Taylor, R. (Eds.), *The Routledge handbook of cognitive linguistics (1st ed.)* (pp. 372-383). Routledge. https://doi.org/10.4324/9781351034708

Suárez-Toste, E. (2007). Metaphor inside the wine cellar: On the ubiquity of personification schemas in winespeak. *Metaphorik*, *12*, 53-64.

Suárez-Toste, E. (2017). Babel of the senses: On the roles of metaphor and synesthesia in wine reviews. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *23*(1), 89-112. https://doi.org/10.1075/term.23.1.04sua

Tay, D. (2015). Metaphor in case study articles on Chinese university counseling service websites. *Chinese Language & Discourse*, *6*, 28-56. https://doi.org/10.1075/cld.6.1.02tay

Ţenescu, A. (2014). The organicist-animist metaphor in Italian wine media discourse. *Social Sciences and Education Research Review*, *1*(2), 62-72.

Ullmann, S. (1957). *The principles of semantics* (2nd ed.). Blackwell.

Wang, V. X., Chen, X., Quan, S., & Huang, C.-R. (2020). A parallel corpus-driven approach to bilingual oenology term banks: How culture differences influence wine tasting terms. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 318-328). Association for Computational Linguistics.

Williams, J. M. (1976). Synaesthetic adjectives: A possible law of semantic change. *Language*, *52*(2), 461-478. https://doi.org/10.2307/412571

Winter, B. (2016). Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience*, *31*(8), 975-988. https://doi.org/10.1080/23273798.2016.1193619

Xu, S. (1963). *Shuowenjiezi [Explaining graphs and analyzing characters]*. Zhonghua Book Company.

Yu, N. (2003). Synesthetic metaphor: a cognitive perspective. *Journal of Literary Semantics*, *32*(1), 19-34. https://doi.org/10.1515/jlse.2003.001

Zawisławska, M., & Falkowska, M. (2019). Metaphors in Polish wine discourse: A corpus approach. *Poznan Studies in Contemporary Linguistics*, *55*, 601-629. https://doi.org/10.1515/psicl-2019-0022

Zeng, W. H., Burgers, C., & Ahrens, K. (2021). Framing metaphor use over time: 'Free Economy' metaphors in Hong Kong political discourse (1997–2017). *Lingua*, *252*, 1-16. https://doi.org/10.1016/j.lingua.2020.102955

Zhao, Q., Huang, C.-R., & Ahrens, K. (2019a). Directionality of linguistic synesthesia in Mandarin: A corpus-based study. *Lingua*, *232*, 102744. https://doi.org/10.1016/j.lingua.2019.102744

Zhao, Q., Xiong, J., & Huang, C.-R. (2019b). Tonggan, yinyu yu renzhi: Tonggan xianxiang zai Hanyu zhong de xitongxing biaoxian yu yuyanxue jiazhi [Synaesthesia, metaphor and cognition: Systematic representations of synaesthesia in Chinese and its linguistic values]. *Zhongguo Yuwen [Studies of the Chinese Language]*, *2*, 240-253.

Zhong, Y., & Huang, C.-R. (2018). Pleasing to the mouth or pleasant personality: a corpus-based study of conceptualization of desserts in online Chinese food reviews. 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC-32): 25th Joint Workshop on Linguistics and Language Processing (JWLLP-25), Hong Kong.

Zhong, Y., & Huang, C.-R. (2020). Sweetness or mouthfeel: A corpus-based study of the conceptualization of taste. *Linguistic Research*, *37*(3), 359-387. https://doi.org/10.17250/khisli.37.3.202012.001

Zhong, Y., Huang, C.-R., & Dong, S. (2022a). Bodily sensation and embodiment: A corpus-based study of gustatory vocabulary in Mandarin Chinese. *Journal of Chinese Linguistics*, *50*(1), 196-230. https://doi.org/10.1353/jcl.2017.0102

Zhong, Y., Wan, M., Ahrens, K., & Huang, C.-R. (2022b). Sensorimotor norms for Chinese nouns and their relationship with orthographic and semantic variables. *Language, Cognition and Neuroscience*. https://doi.org/10.1080/23273798.2022.2035416

Contextual-Boosted Deep Neural Collaborative Filtering Approach for Arabic Textual Documents Recommendation

Ons Meddeb[1,2] **Mohsen Maraoui**[2] **Mounir Zrigui**[2]

[1]University of Sousse, Higher Institute of Computer Science and Communication Techniques ISITCom, Hammam Sousse 4011, Tunisia

[2]University of Monastir, Research Laboratory in Algebra, Numbers Theory and Intelligent Systems RLANTIS, Monastir 5000, Tunisia

meddeb.ons@gmail.com, maraoui.mohsen@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract

The technological advancement and the expansion of using the internet has made it possible to communicate between people and machines. Many exchanged textual resources have been given simultaneously make users have difficulty in choosing the most appropriate items that will improve their knowledge. To deal with data sparsity problem, review-based recommendation systems have shown potential in a wide range of Natural Language Processing (NLP) tasks. Due to the high-dimensionality and the complex semantics of Arabic textual data, review-based approach is proposed for documents recommendation. Two parallel neural networks are introduced to learn items properties and users' behaviors. Indeed, contextualized word representation model is used. Then, Gated Recurrent Units (GRU) is attached for extracting high-level-semantic features. After that, a shared layer modeled complex interactions between the latent vectors of users and items to improve subsequently rating prediction. Experiments showed the superiority of our proposed model compared with state-of-the-art methods.

1 Introduction

Many technological advances have made it possible to communicate between peoples and machines (Slimi et al., 2022). The greatest benefit of delivering resources provided users with an opportunity to try a new learning style and explore its advantages (Hazar et al., 2022). With the increasing abundance of data over the internet, research on high-quality recommendation systems has gained interest in industry and academic. It has become very important for online platforms (e.g. e-commerce, music, books, social media, advertising, etc.) to provide the users what they need and like without wasting time searching (Dhelim et al., 2022).

Most early recommendation systems use Collaborative Filtering methods (CF) that focus on learning accurate representations of user preferences and item features. Previous models used numeric ratings given by users as inputs. However, they suffered from data sparsity problem (Duan et al., 2022). They had difficulties for learning representations and generating reliable recommendations for users or items with few ratings. Another drawback of CF is that it did not make full use of the available context information like item attributes or user profile to make thereafter recommendations. To alleviate these problems, the use of textual reviews has attracted growing attention. Users can explain their opinions underlying their given ratings. They contain valuable and rich information that cannot be obtained from ratings alone (Jian et al., 2022).

Most of reviews-based models have been focused on English language but not on Arabic. This is due to the richness of Arabic specificities (Meddeb et al., 2021a): a review can include different forms in the vocabulary or syntactic and semantic representations. For example, a word can have more than one lexical category in different contexts what changes the meaning of the sentence (Mahmoud and Zrigui, 2017). In this context, reviews-based recommendation models have attracted attention in different Natural Language Processing applications (e.g., product recommendation, information retrieval, social networks, etc.).

In this paper, contextual-boosted deep neural CF approach is proposed for Arabic textual documents recommendation. This model learns user and item representations simultaneously using two parallel networks. Each one is based on Arabic Bidirectional Encoder Representations from Transformers (AraBERT) for textual reviews embedding. It is efficient to extract meaningful features adaptable to arbitrary contexts that cannot be extracted from traditional word embedding like word2vec and GloVe. For more interpretability, Gated Recurrent Units (GRU) architecture is applied for modeling more semantics from reviews. Once user and item representations are learned, they are concatenated together in a shared hidden space and finally fed to multilayer perceptron (MLP) that is used as an interaction function for rating prediction.

This paper is organized as follows: Section 2 presents a literature review. Section 3 and 4 detail the components and experiments of the proposed approach. Finally, section 5 describes conclusions and future work.

2 Literature Review

Recommendation systems have become widely used for alleviating the overload of information available online. For example, videos recommendation (YouTube), films recommendation (Netflix), Music recommendation (Last.fm), and Books recommendation (Goodreads, Amazon, etc.) (Meddeb et al., 2021b). In this section, the related works to our research are presented.

2.1 CF-based recommendation

Collaborative filtering (CF) is a dominant state-of-the-art recommendation method in which peoples how share similar preferences in the past tend to have similar choices in the future (Meddeb et al., 2021c). The most successful CF methods were based on Matrix Factorization (MF). Using historical records (e.g., ratings, clicks, consumptions, etc.), their main idea is to construct an implicit semantic model representing users and items as vectors of latent factors (called embedding); and modeling thereafter user-item interactions using the inner product operation. Several recommendation methods have been employed MF models.

For more interpretability, Lee and Seung (2001) introduced Non-Negative Matrix Factorization model (NGMF), in which a non-negativity constraint was proposed based on Singular Value Decomposition (SVD). Such optimal estimation methods have demonstrated the existence of overfitting problem due to the sparsity of user-item interactions. To deal with this issue, probabilistic factor models have been proposed. Mnih and Salakhtdinov (2007) proposed Probabilistic Matrix Factorization (PMF) method. It scaled linearly with the number of observations and performed well. In the same idea, Fang et al. (2020) proposed a Bayesian Latent Factor Model (BLFM). Based on the observed user-item interactions, they introduced a constraint on latent factor, and established a likelihood function. However, the majority of MF based methods were sub-optimal for learning rich real world data and complicated user-item interactions. To address these issues, neural networks have been integrated into recommender architectures. He et al. (2017) proposed Neural Collaborative model (NCF). They replaced the inner product with Multilayer Perceptron (MLP) model. It was useful for learning meaningful user-item interactions granting a high degree of nonlinearity and flexibility. Similarly, Bai et al. (2017) proposed a Neighborhood-based NCF (NNCF), in which an MLP layer was placed above the concatenated user-item embedding. By cons, Zhang et al. (2017) and Wang et al. (2017) placed it above the element-wise product of user and item latent vectors. Recently, Chen et al. (2019) proposed a Joint NCF (J-NCF) approach. They adopted the rating to explore both user and item features. MLP layers were applied for extracting latent vectors of users and items and modeling thereafter the interactions between them.

Although CF techniques have shown good performances for many applications, the sparsity problem is considered as one of their significant challenges. It is not easy for them to recommend items with few ratings. To alleviate these issues, users' textual reviews are used as auxiliary information.

2.2 Deep learning-based review modeling

Textual reviews contain rich semantic information about users and items, in which opinion written by users can reveal some information on rating behavior, and also opinions written for items may contain indications on their features (Zheng et al., 2017). For instance, McAuley and Leskovec (2013) proposed Hidden Factors and Topics (HFT) model for understanding rating dimensions with review text. User reviews were modeled using matrix factorization. Indeed, the topic distribution of each review was produced by the latent factors of the corresponding item. In the same vein, King (2014) proposed a unified model called Ratings Meet Reviews (RMR). Content-based filtering with CF incorporated both ratings and reviews. Topic modeling techniques were applied to the review. The obtained topics were aligned with rating dimension to improve rating prediction. Bao et al. (2014) applied Latent Dirichlet Allocation (LDA) and NGMF for latent vectors extraction from ratings and reviews. Then, transform function predicted missing ratings. Saeed et al. (2021) demonstrated the superiority of LDA than Principal Component Analysis (PCA) for Arabic reviews modelling and sentiment classification.

Recently, deep learning modeled efficiently auxiliary review information, such as the textual descriptions of items and preferences of users. They have good capabilities in modeling semantic information by considering the context of words from reviews. The first deep model was proposed by Zheng et al. (2017). Two parallel Convolutional Neural Networks (CNN) represented items' features and users' behaviors from reviews. The use of CNN was helpful in transforming features to high-level abstraction from a stacked of convolution followed by pooling layers (Bellagha and Zrigui, 2021). A shared layer fused them and a FM captured user-item interactions. After that, Chen et al. (2018) proposed NARRE model similar to DeepCoNN, in which two parallel CNNs were used for users and items modeling. Rather than concatenating reviews to one long sequence the same way that DeepCoNN does, an attention mechanism learned the review usefulness. The obtained attention weights are integrated into user and item representations to enhance the embedding quality and the subsequent prediction accuracy. Both DeepCoNN and NARRE employed traditional word embedding.

Everyone has an imposed background from his mother tongue and will have his own difficulties wide different from any another one speaking another language (Trigui et al., 2022). Following the literature, the majority of previous recommender methods have been focused on English resources. In contrast, few works have been concentrated on Arabic language. It is one of the poorly endowed languages that must be treated specifically (Haffar et al., 2021):

- *Complex morphological language:* the existence of diacritics and stacked letters. Arabic is an inflectional, derivational and non-concatenative language. (Mahmoud et al., 2021a).
- *Agglutinative language:* lexical units of words vary in number and in bending according to the grammatical relationships within sentences. (Mahmoud et al., 2021b)
- *Ambiguous language:* words can have more than one sense that is dependent on the context of use (Meddeb et al., 2017). For example, the word "قبل" can be a verb (accept) or an adverb (before) or a noun in its plural form (kisses). (Sghaier and Zrigui, 2020)

3 Proposed Approach

This section details the proposed approach for Arabic textual documents recommendation. The overall model is illustrated in Fig. 1. It has two parallel networks to model user and item embedding that are thereafter concatenated for modeling user-item interactions and rating prediction.

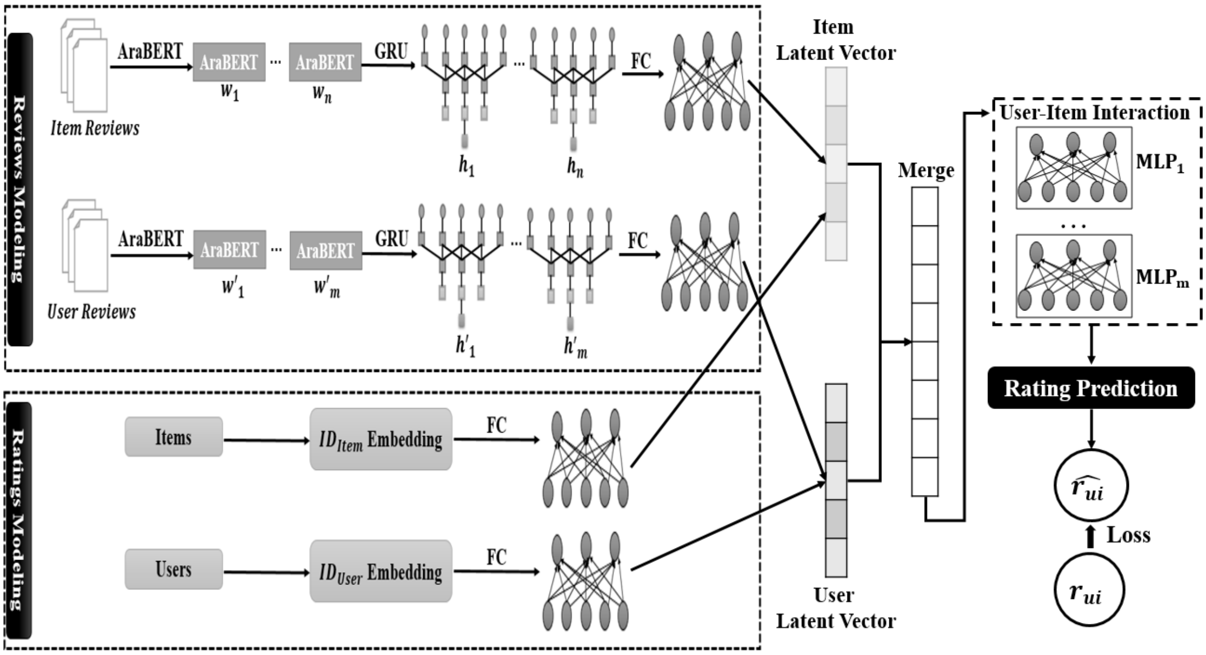


Fig. 1. Proposed architecture.

3.1 Rating Modeling Phase

Suppose that there are M users and N items denoted as $U = \{user_1, user_2, \ldots, user_M\}$ and $I = \{item_1, item_2, \ldots, item_N\}$. The user-item rating matrix R of dimension $N \times M$ is composed by the rating $r_{u,i}$ given by a user $user_u$ to an item $item_i$. For rating modelling, lookup function $\emptyset$ is applied to project the sparse representations into dense vectors using as inputs the identities of users $\{ID_1, ID_2, \ldots, ID_u\}$ and items $\{ID'_1, ID'_2, \ldots, ID'_i\}$, as follows in Eq. (1-2):

$$V^u_{1:M} = \emptyset(ID_1), \emptyset(ID_2), \ldots, \emptyset(ID_M) \qquad (1)$$

$$V^i_{1:N} = \emptyset(ID'_1), \emptyset(ID'_2), \ldots, \emptyset(ID'_N) \qquad (2)$$

Latent vectors of items $I_{Ratings}$ and users $U_{Ratings}$ are created through two fully connected layers according to the numerical ratings, as denoted in Eq. (3-4):

$$U_{Ratings} = ReLU(W.V^u_{1:M} + b_u) \qquad (3)$$

$$I_{Ratings} = ReLU(W'.V^i_{1:N} + b_i) \qquad (4)$$

3.2 Reviews Modeling Phase

Due to the sparseness and high-dimensionality of textual data, there are many difficulties for their NLP such as semantic diversity, metaphor expression and grammatical specificity. To solve these problems, reviews modelling process is proposed based on AraBERT and deep learning for rating prediction and Arabic textual documents recommendation. In this section, we will only illustrate the *user modeling process* because the same is also used for items with their inputs as the only difference. Indeed, each user or item is represented as a feature vector in K-dimensional latent factor space as follows: First, Arabic Bidirectional Encoder Representation from Transformers (AraBERT) represent the text with dynamic word vectors according to the context information. It can be adjusted according to the word meaning while the context information is fused. Then, Gated Recurrent Unit (GRU) extract the contextual features from the text.

AraBERT Representation Layer: Given an input set of user-written reviews $S^{user}_u = \{s_{u1}, s_{u2}, \ldots, s_{uj}\}$ where J is the total number of Arabic reviews from user u. S^{user}_u is fed to a pre-trained model that builds upon the transformer architecture. As our initial embedding model, we use $AraBERT$ of 4 encoder layers and 4 self-attention heads. It has 256 hidden dimensions, that are directly utilized later as the fixed embedding dimension. The parameters are fine-tuned during the training process of our model. Indeed, each review is tokenized into words. Because rating prediction is not a sentence pairing task, AraBERT model takes thereafter each review as a single textual segment composed of 256 tokens. The obtained sequences are then passed through a stack of transformer encoders to obtain their respective contextualized representations as follows in Eq. (5):

$$h_{[CLS],u} = \{h_{[CLS],u1}, h_{[CLS],u2}, \dots, h_{[CLS],uj}\} \quad (5)$$

Where: $h_{[CLS],u} \in R^{j\times 256}$.

In theory, any encoder layer may be selected to provide the hidden state of [CLS] as the review's representations. In this study, we select the 4th layer. Following Sun et al. (2019), they have illustrated that the predictive capability of this layer is the best among the others. The final user embedding $P_u \in R^{1\times 256}$ is generated by calculating the average of the [CLS] representations of the reviews written by a user u, as defined in Eq. (6):

$$p_u = \frac{1}{J}\sum_{t=1}^{J} h_{[CLS],ut} \quad (6)$$

Similarly, the item embedding $Q_i \in R^{1\times 256}$ is generated with the same manner from the item embedding network as defined in Eq. (7):

$$q_i = \frac{1}{J}\sum_{t=1}^{J} h_{[CLS],it} \quad (7)$$

To sum up, each user u 's review set $S_u^{user} = \{s_{u1}, s_{u2}, \dots, s_{uj}\}$ and item i 's review set $S_i^{item} = \{s_{i1}, s_{i2}, \dots, s_{ij}\}$ can be represented as comprehensive vectors $P_u^{user} = \{p_1, p_2, \dots, p_N\} \in R^{N\times 256}$ and $Q_i^{item} = \{q_1, q_2, \dots, q_m\} \in R^{M\times 256}$, respectively, where N and M are the total numbers of users and items.

Fine-Tuning Layer: The fine-tuning layer is proposed to focus on the effective information in the reviews used for recommendation. Following the literature, several recent NLP studies showed the relevance of RNN models for modeling sequential data and extracting hidden contextual states. Gated Recurrent Units (GRU) architecture is a different version of RNN widely used because of its time efficiency characteristics. Instead of LSTM-RNN, it has only two gates:

- The reset gate r_t jointly controls the calculation from the previous hidden state h_{t-1} to the current one h_t.
- The update gate z_t controls both the current input data and the previous memory information h_{t-1}. It determines how much h_{t-1} is passed to the next state.

GRU is characterized by fewer parameters, that can affect the same effect as LSTM while reducing the training time. Another advantage of using GRU is that its hidden layer uses long short-term memory and gated recurrent units to hold the long-term dependencies, which are inherent in the text regardless of lengths and occurrences. The specific gate unit is calculated as follows (Eq. 8-11):

$$z_t = ReLU(W_z.[h_{t-1}, x_t]) \quad (8)$$

$$r_t = ReLU(W_r.[h_{t-1}, x_t]) \quad (9)$$

$$\tilde{h}_t = tanh(W.[r_t \times h_{t-1}, x_t]) \quad (10)$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \quad (11)$$

Where: x_t is the input of GRU model and obtained by AraBERT pre-training language model; $W \in R^{m\times n}$ are the weight matrices of reset r_t and update z_t gates, n is the number of the hidden units; h_{t-1} is the previous hidden state; $\tilde{h}_t$ is the candidate hidden state; and the operator $\cdot$ denotes an element point-wise multiplication.

The output of the GRU model is transmitted to a fully connected layer to model the final latent vectors of reviews written by users $U_{Reviews}$ and those written for items $I_{Reviews}$ as follows in Eq. (12-13):

$$U_{Reviews} = ReLU(W_u.h_t^{(l)} + b_u) \quad (12)$$

$$I_{Reviews} = ReLU(W_i.h_t^{(l)} + b_i) \quad (13)$$

Where: $W \in R^{f\times k}$ is the weight matrix, $b \in R^k$ is the bias term, ReLU is the Rectified Linear Unit activation function.

3.3 Heterogeneous Information Merge Phase

The final representations of users' preferences included the latent vectors of user-item ratings and user reviews, as defined in Eq. (14):

$$F_u = [U_{Ratings}, U_{Reviews}] \quad (14)$$

The final representation of items' features is composed by the latent vectors of item reviews and ratings, as defined in Eq. (15):

$$F_i = [I_{Ratings}, I_{Reviews}] \quad (15)$$

3.4 Rating Prediction Phase

User-Item Interaction Modeling: The final feature vectors of the users $F_u \in R^D$ and items $F_i \in R^D$ are concatenated by applying the dot product operation as defined in Eq. (16). For learning the interactions between user and item representations and modelling the CF effect, an affinity score $\widehat{rate}_{u,i}$ is defined as user u 's preference for item i as denoted in Eq. (19). To do this, Multi-Layer Perceptron (MLP) model is applied on top of the concatenated user-item embedding to provide further flexibility and non-linearity as defined in Eq. (17-19).

$$h_o = F_u \odot F_i \tag{16}$$

$$h_1 = ReLU(W_1.h_o + b_1) \tag{17}$$

$$\ldots$$

$$h_L = ReLU(W_L.h_{L-1} + b_L) \tag{18}$$

$$\widehat{rate}_{u,i} = W_{L+1}h_L + b_L \tag{19}$$

Where : h_0 is the concatenated user-item embedding in the shared hidden space; h_L is the Lth MLP layer, W_L is the weight matrix, b_L is the bias vector, ReLU is the activation function and $\widehat{rate}_{u,i}$ denotes the predicted rating that user u gives to item i . In our model, three layered MLP were useful to learn efficiently more abstractive features.

Learning: For training, the Mean Squared Error (MSE) is used as the loss function. To optimize the performance of our model, MSE is minimized between the output score $\widehat{rate}_{u,i}$ from our model and the real score $rate_{u,i}$. It is defined as follows in Eq. (20):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(rate_{u,i} - \widehat{rate}_{u,i})^2 \tag{20}$$

Where: N refers to the training samples, $rate_{u,i}$ is the ground-truth rating given by user u to item i, and $\widehat{rate}_{u,i}$ is the predicted rating.

4 Experiments

4.1 Dataset

The experiments are conducted on Books Reviews in Arabic Dataset BRAD1.0. It consists of 4,993 Arabic books of different genres (historic, religion, health…) collected from the Goodreads web site; 76,530 users and 510,600 user-item interactions composed by reviews and ratings in the range of [1,5], as illustrated in the following Table 1:

Number of users	Number of books	Number of user-item interactions	
		Number of ratings	Number of reviews
76,530	4,993	510,600	510,600

Table 1: BRAD1.0 dataset

4.2 Parameters settings

This section represents the parameters of our proposed models. Table 2 depicts the configurations of AraBERT[1], GRU and MLP models, for textual reviews embedding and user-item interaction modeling:

Model	Parameter	Value
AraBERT	Hidden layers	4
	Attention heads	4
	Hidden size	256
	Dropout	0.1
	Optimizer	Adam
	Epochs	20
	Activation function	ReLU
GRU	Hidden units	256, 128, 64
	Activation function	ReLU
	Epochs	20
	Batch size	128
	Dropout rate	0.2
MLP	Latent factors	16
	Activation function	ReLU
	Hidden layers	128, 64, 8
	Batch size	256
	Droput rate	0.5

Table 2: Parameters configurations

4.3 Evaluation Metric

Root Mean Squared Error (RMSE) is used as an evaluation metric. It is the most used for rating prediction and recommendation systems.

Given N user-item ratings, the RMSE score is defined as follows in Eq. (21):

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(rate_{u,i} - \widehat{rate}_{u,i})^2} \tag{21}$$

4.4 Baselines

As depicted in Table 3, we compare our model with several competitive baselines, including CF- and deep learning-based methods, using reviews and ratings:

- **PMF (Probabilistic Matrix Factorization):** is based on a Gaussian distribution to model the latent factors for users and items.
- **SVD ++ (Singular Value Decomposition):** extends SVD with neighborhood method. The item-item similarity using a novel set of item factors.
- **J-NCF (Joint-Neural Collaborative Filtering):** coupled deep features learning and deep interactions modeling with a rating matrix. This was due by using MLP model.
- **DeepCoNN (Deep Cooperative Neural Network):** utilizes deep learning to

[1] https://github.com/aub-mind/arabert/tree/master/arabert

jointly model user and item from textual reviews.

Models	Features		
	Ratings	Reviews	Deep learning
PMF	×	-	-
SVD++	×	-	-
J-NCF	×	-	×
DeepCoNN	-	×	×
Our model	×	×	×

Table 3: Features of state-of-the-art methods

4.5 Discussion

Using BRAD1.0 dataset, several observations can be made from the rating prediction results of our model compared to the state-of-the-art methods as illustrated in Table 4:

Methods	RMSE
PMF	1.355
SVD++	1.215
J-NCF	1.050
DeepCoNN	0.905
Our model	0.855

Table 4: Experimental results of baselines

As illustrated in J-NCF model, user-item interactions modeled efficiently in a non-linear way, which is the limitation of traditional CF methods based on the dot product. It achieved 1.050 RMSE higher than PMF (1.355 RMSE) and SVD++ (1.215 RMSE).

Reviews-based methods like DeepCoNN performed better than classic CF models (PMF, SVD and J-NCF) that only consider the rating matrix as the input. It obtained 0.905 RMSE. This demonstrated the usefulness of review information. It was complementary to ratings specially to improve the representation quality of latent factors of user preferences and item features, and thereafter rating perdition.

Although review information was useful in recommendation, the performance could vary depending on how it was modeled. Indeed, the application of CNN was efficient for modeling reviews and extracting latent vectors of users and items.

Although that it did not need to manually label features, their performance was not satisfactory for analyzing the specificities of an ambiguous language like Arabic. To deal with this, we proposed a contextual-boosted deep neural collaborative filtering approach for Arabic textual documents recommendation. As shown in Fig.2, our proposed approach consistently outperformed all the cited baseline methods achieving 0.855 RMSE for the following reasons:

The model used the advanced AraBERT language model. It modeled efficiently deep semantics of words and dynamically generated high-quality of word vectors unlike traditional language models (e.g., GloVe and word2vec).

Then, GRU was useful to solve the related dependencies of long sentences by considering the context information.

Finally, experiments demonstrated the superiority of our model regarding the dimensionality of the user and item vectors and the hidden layers' number of MLP while fixing the other parameters: the width of embedding layer (K=16) and the number of hidden layers (3-MLP) improved the recommendation performance achieving 0.855 RMSE.

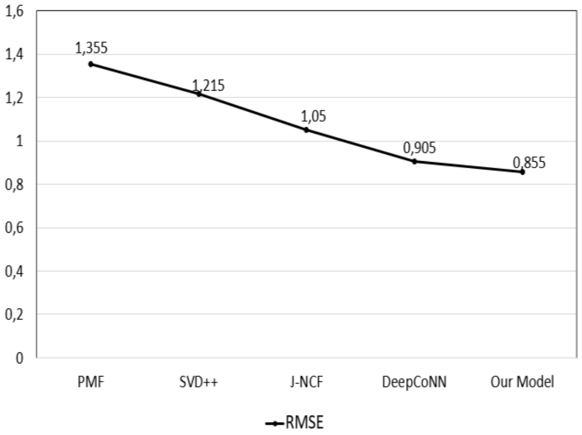


Fig. 2. State-of-the-art comparisons according to RMSE scores

5 Conclusion and Future Work

A contextual-boosted deep neural collaborative filtering approach for Arabic textual documents recommendation is proposed. It combined ratings and textual reviews seeing their complementary to improve the representation quality of latent factors of user preferences and item features.

The pre-trained AraBERT language model was useful to mine the deep semantics of Arabic words and dynamically generate high-quality vectors. Then, GRU modeled efficiently the related dependencies of long sentences by considering the context information. After that, MLP helped to model complex user-item interactions and improve rating prediction. Our proposed model is validated by experiments realized on BRAD1.0 dataset. They consistently outperformed other state-of-the-art methods namely PMF, SVD++ and J-NCF.

In linguistics, both of the previous and the succeeding words influence the current word semantic. For this reason, bidirectional recurrent neural network will be adopted in the future to learn long term dependencies in both forward and backward directions to improve the user-item textual information modeling process.

References

Bai T., Wen J., Zhang J., and Zhao W. X. 2017. A neural collaborative filtering model with interaction-based neighborhood, ACM on Conference on Information and Knowledge Management (ICKM): 1979–1982.

Bao Y., Fang H., Zhang J. 2014. TopicMF: simultaneously exploiting ratings and reviews for recommendation, 28th AAAI Conference on Artificial Intelligence, Quebec, Canada: 2-8.

Bellagha M. L., and Zrigui M. 2020. Speaker naming in tv programs based on speaker role recognition, IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), Antalya, Turkey: 1-8.

Chen C., Zhang M., Liu Y., and Ma S. 2018. Neural attentional rating regression with review-level explanations, International World Wide Web Conferences Steering Committee (WWW): 1583–1592.

Chen W., Cai F., Chen H., and Rijke M. D. 2019. Joint neural collaborative filtering for recommender systems, ACM Transactions on Information Systems, 1(1): 1-30.

Devlin J., Chang M., Lee K., and Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dhelim S., Aung N., Bouras M., Ning H., and Cambria E. 2022. A survey on personality-aware recommendation systems, Artificial Intelligence Review, volume 55: 2409-2454.

Duan R., Jiang C., and Jain H. 2022. Combining review-based collaborative filtering and matrix factorization: A solution to rating's sparsity problem, Decision Support Systems, volume 156:113748.

Fang J., Zhang X., Hu Y., Xu Y., Yang M., and Liu J. 2020. Probabilistic latent factor model for collaborative filtering with Bayesian inference, 25th International Conference on Pattern Recognition (ICPR), Milan, Italy: 73-80.

Haffar N., Ayadi R., Hkiri E., and Zrigui M. 2021. Temporal ordering of events via deep neural networks, 16th International Conference on Document Analysis and Recognition (ICDAR), Lausanne, Switzerland: 762-777.

Hazar M. J., Maraoui M., and Zrigui M. 2022. Recommendation system based on video processing in an E-learning platform, Journal of Human University (Natural Sciences), 49(6): 49-61.

He X., Liao L., Zhang H., Nie L., and Chua T.S. 2017. Neural collaborative filtering, 26th International Conference on World Wide Web (WWW), Perth Australia: 173–182

Jian P. K., Patel A., Kuari S., and Pamula R. 2022. Predicting airline customers' recommendations using qualitative and quantitative contents of online reviews, Multimedia and Tools, volume 81: 6979-6994.

King G. L. L. 2014. Ratings meet reviews, a combined approach to recommend, 8th ACM Conference on Recommender Systems (RecSys), Silicon Valley, CA, USA: 105–112.

Lee D., and Seung H. S. 2001. Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada: 556–562.

Mahmoud A., and Zrigui M. 2017. Semantic similarity analysis for paraphrase identification in Arabic texts, 31st Pacific Asia Conference on Language, Information and Computation (PACLIC), Philippine: 274-281.

Mahmoud A., and Zrigui M. 2021a. BLSTM-API: Bi-LSTM recurrent neural-based approach for Arabic paraphrase identification, Arabian for Science and Engineering, volume 46: 4163-4174.

Mahmoud A., and Zrigui M. 2021b. Semantic similarity analysis for corpus development and paraphrase detection in Arabic, International Arab Journal of Information Technology (IAJIT), volume 18: 1-7.

McAuley J., and Leskovec J. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text, 7th ACM Conference on Recommender systems, Hong Kong, China: 165–172.

Meddeb O., Maraoui M., Aljawarneh S. 2017. Hybrid modelling of an off line Arabic handwriting recognition system: results and evaluation, International Journal Intelligent Enterprise, 4(½): 168–189.

Meddeb O., Maraoui M., and Zrigui M. 2021a. Deep learning based semantic approach for Arabic textual documents recommendation, IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey: 1-6.

Meddeb O., Maraoui M., and Zrigui M. 2021b. Arabic text documents recommendation using joint deep representations learning, Procedia Computer Science, 192(1): 812-821.

Meddeb O., Maraoui M., and Zrigui M. 2021c. Personalized smart learning recommendation system for Arabic users in smart campus, International Journal of Web-based learning and Teaching Technologies, 51(5): 1734-1747.

Mnih A., and Salakhutdinov R. 2007. Probabilistic matrix factorization, Advances in Neural Information Processing Systems, volume 20: 1257-1264.

Saeed R. M. K., Rady S., and Gharib T. F. 2021. Optimizing sentiment classification for Arabic opinion texts, Cognitive Computation, volume 13: 164–178.

Sghaier M. A., and Zrigui M. 2020. Rule-based machine translation from Tunisian dialect to modern Arabic standard, Procedia Computer Science, volume 196: 310-319

Slimi A., Zrigui M., and Nicolas H. 2022. MuLER: Multiplet-loss for emotion recognition, International Conference on Multimedia Retrieval (ICMR), Newark, NJ, USA: 435-442.

Sun C., Qiu X., Xu Y., and Huang X. 2019. How to fine-tune bert for text classification?: arXiv preprint arXiv:1905.05583.

Trigui A., Mars A., Ben Jannet M. A., Maroui M., and Zrigui M. 2022. Foreign accent classification for Arabic speech learning: 1-5.

Wang X., He X., Nie L., and Chua T. 2017. Item silk road: Recommending items from information domains to social users, 40th International ACM SIGIR Conference on Research and Development in Information Retrieval: 185–194.

Zhang Y., Ai Q., Chen X., and Croft W. 2017. Joint representation learning for topn recommendation with heterogeneous information sources. ACM on Conference on Information and Knowledge Management (ICKM): 1449–1458.

Zheng L., Noroozi V., and Yu P. S. 2017. Joint deep modeling of users and items using reviews for recommendation, 10th ACM International Conference on Web Search and Data Mining (WSDM): 425-434.

Learning Sentence Embeddings in the Legal Domain with Low Resource Settings

Sahan Jayasinghe[*,1], **Lakith Rambukkanage**[*], **Ashan Silva**[*], **Nisansa de Silva**[*],
Amal Shehan Perera[*], and **Madhavi Perera**[**]

[*]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[**]Parliament of Sri Lanka, Sri Lanka
[1]sahanjayasinghe.17@cse.mrt.ac.lk

Abstract

As Natural Language Processing is evolving rapidly, it is used to analyze domain specific large text corpora. Applying Natural Language Processing in a domain with uncommon vocabulary and unique semantics requires techniques specifically designed for that domain. The legal domain is such an area with unique vocabulary and semantic interpretations. In this paper we have conducted research to develop sentence embeddings, specifically for the legal domain, to address the domain needs. We have carried this research under two approaches. Due to the availability of a large corpus of raw court case documents, an Auto-Encoder model which re-constructs the input sentence is trained in a self-supervised approach. Pretrained word embeddings on general corpora and word embeddings specifically trained on legal corpora are also incorporated within the Auto-Encoder. As the next approach we have designed a multitask model with noise discrimination and Semantic Textual Similarity tasks. It is expected that these embeddings and gained insights would help vectorize legal domain corpora, enabling further application of Machine Learning in the legal domain.

1 Introduction

Natural Language Processing (NLP) is advancing rapidly in the research domain as well as in the practical applications. Several researches have been conducted in the recent past, that have made ground breaking progress but some are yet to be discovered and applied in practical applications. It is not trivial to apply ML techniques directly on unstructured data and NLP approaches different aspects of these problems. Also the advantages of using NLP is best utilized in fields which handle large amounts of textual data. The Legal Domain is such a domain where an abundance of textual data is available, and legal corpora is growing on a daily basis.

1.1 Case Law

Case Law documents are one of the aspects which contributes to the rapidly growing textual data in the legal domain. In case law, records of past cases with their evidence arguments and judgment are kept in order to be used as reference and grounds for ongoing cases (cor, 2020). The usage of similar cases with respect to the current case as grounds, is why these documents are very important in a predictive sense. They serve as a good training data source for researches that explore the application of NLP in the Legal Domain.

1.2 Word and Sentence Embeddings

NLP consists of many techniques such as parts of speech tagging, sentiment analysis, text generation and language translation among many others. Regardless, unstructured data requires numerical representation to be analysed using ML techniques. For this transformation, often times, the text data is converted in to vector format or in other words, embeddings in order to be processed using machine learning and deep learning techniques. There are a lot of state of the art word embeddings such as Word2Vec (Mikolov et al., 2013a), Glove (Pennington et al., 2014), FastText(Mikolov et al., 2018), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019) and sentence embeddings available today such as Sentence-BERT (SBERT) (Reimers and Gurevych, 2019),Universal Sentence Encoder (USE) (Cer et al., 2018) and InferSent (Conneau et al., 2017). The main drawback in directly using these embeddings and approaches is that they have been designed and evaluated for general purpose datasets and applications. In addition the datasets used in these approaches are general purpose corpora. These embeddings are very useful for domain independent tasks but may perform poorly in domain dependant tasks.

1.3 Domain specific embeddings

The legal domain has rare vocabulary terms such as *"Habeas Corpus"*, that are rarely found in domain independent corpora. In addition, some common words imply a context specific meaning in the legal domain. For example the word *"Corpus"* in *"Habeas Corpus"* and *"Text Corpus"* gives different meanings in the legal domain. This aspect is also not captured when training with general corpora. Also approaches that are specifically designed for the legal domain should be researched to address the inherent complexities of the domain. Considering all these facts, this paper discusses the designing and training of a legal domain specific sentence embedding based on criminal sentence corpora.

2 Related Work

A lot of ground-breaking researches have been conducted in the past years, contributing to the evolution of NLP. This research makes use of a lot of these researches for both intuition and auxiliary purposes, which are discussed in this section. It is important to highlight that most of the sentence embeddings that have demonstrated state of the art performances, have been trained with large annotated datasets as discussed in subsection2.2.

2.1 Word Embeddings

The targeted word embedding of Word2Vec (Mikolov et al., 2013a) is in the continuous vector format. They have considered the facts, that Latent Semantic Analysis (LSA) is poor at preserving linear regularities of embeddings, and the computational demand of Latent Dirichlet Allocation (LDA) with large datasets. The Word2Vec architecture is a 2 layer neural network which uses two techniques, 1) Continuous Bag of Words (CBOW) and 2) Skip-gram. Words that appear in a similar context is assumed to have a similar meaning. CBOW is preferred for small corpora and faster training whereas Skip-gram performs better with large corpus but trains slower. Also in a later publication (Mikolov et al., 2013b) they propose several improvements. One improvement is the sub sampling of frequent words which reduces the training time preserves rare words and increases their accuracy. Also, negative sampling is introduced where set of words with incorrect label is used. The difference of phrases in contrast to individual meanings of words, is also addressed by allowing them to be individual tokens.

Since the emergence of multi-headed attention (Vaswani et al., 2017) with transformer architectures, BERT (Devlin et al., 2018) came up with language modeling techniques to generate word embeddings. BERT is designed in a way that it can be fine tuned for a specific task with minimum changes to the architecture, unlike the other word embeddings. Since BERT uses sub words, out of vocabulary words can be also embedded easily which is an important aspect, but they may not be as accurate when originally trained on.

As an advancement to the Masked Language Modeling (MLM) proposed by BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) introduces many improvements. They have experimented on the impact of tuning several hyper-parameters and optimization algorithms as well as the training data. They have identified the Next Sentence Prediction (NSP) task is not improving accuracy and have only used MLM. Also in contrast to static masking in BERT (Devlin et al., 2018) they have identified dynamic masking improves accuracy. Finally by increasing batch size, adding training data and pre-training longer have achieved better performance than existing BERT, XLnet (Yang et al., 2019) models in many evaluation tasks.

XLnet (Yang et al., 2019) is developed with the intention of benefiting from both aspects of 1) Auto-regressive models that use Long Short Term Memory (LSTM) and 2)Auto-encoding models such as in BERT (Devlin et al., 2018). Also it is designed to be used for mainstream NLP tasks. In order to combine the two approaches they use the auto-regressive nature of referring to only the context seen before, and use a generated permutation of the input to give access to the whole context. With these improvements they have been able to beat BERT (Devlin et al., 2018) at many mainstream NLP evaluations.

Glove (Pennington et al., 2014) takes into consideration the drawbacks of the existing model families 1) Matrix factorization models and 2) Context window related models. The context used in Glove is derived from a window where the approximation of similarity for the two word pairs considered at a time, is inversely proportionate to the distance between the two words. Glove has been trained on comparatively a large amount of data than other existing methods but uses fewer dimensions, and has been able to beat the performance of them at many evaluation tasks.

FastText (Mikolov et al., 2018) is designed to account for morphology of words, where words can take different forms which is not captured in models like Word2Vec(Mikolov et al., 2013a). FastText does this by using N-grams as the tokens, which leaves provision for out of vocabulary or misspelled words. They have achieved better accuracy with significant drop in training time.

2.2 Sentence Embeddings

SBERT (Reimers and Gurevych, 2019) uses two BERT (Devlin et al., 2018) encoders to encode the words in two sentences. The vectors of the words are then pooled using mean pooling to get a singe vector for each sentence. These vectors can then be passed on to a Soft-max classifier for classification or cosine similarity function for regression. The Stanford Natural Language Inference (SNLI) dataset is used to train this model.

Universal Sentence Encoder (USE) (Cer et al., 2018) describes two approaches for sentence embeddings, 1) transformer base models and 2) Deep Averaging Networks (DAN). The transformer based approach is high in complexity and consumes more resources but it is more accurate. In contrast DAN based models give less accuracy with less resource consumption. Multiple downstream tasks are used to make the model more generalized. Out of the two approaches, they have concluded that overall, the transformer based approach is better in accuracy.

Unlike the common unsupervised approaches, InferSent (Conneau et al., 2017) model is a sentence embedding trained with a supervised approach. Initially they have experimented with several architectures with techniques such as LSTMs and Gated Recurrent Units (GRU) and different pooling techniques. Since the approach is supervised, researchers have used the SNLI dataset. They have demonstrated that using Bidirectional LSTM (Bi-LSTM) and max pooling along with a supervised approach can outperform existing unsupervised approaches.

Researchers (Wieting et al., 2015), have formalized a way to obtain sentence embeddings considering the paraphrastic nature of sentence pairs by calculating cosine similarities of embeddings. Initially they have experimented with many approaches, with the simplest being averaging word vectors, to LSTMs. They have identified that averaging word vectors is outperforming LSTMs at many tasks, but LSTMs are better at sentiment classification tasks.

The requirement for a baseline to evaluate sentence embeddings, is addressed by researchers (Arora et al., 2017) which is mainly motivated by the work of (Wieting et al., 2015) They have stated that it can be used to evaluate domain specific sentence embeddings. Compared to the (Wieting et al., 2015) approach, they have identified that, rather than using a simple averaging function, smoothing inverse frequency techniques perform better, even more so than some LSTM and RNN approaches. Similar to the research (Wieting et al., 2015) they have identified LSTMs and RNNs are much capable of sentiment related tasks.

Sent2Vec (Pagliardini et al., 2018) have identified that all language representational learning approaches have either complex deep models trained on large datasets with expensive computing or Matrix factorization methods which is less computationally expensive but effective on large corpora. Similar to ”Towards universal paraphrastic sentence embeddings” (Wieting et al., 2015), researches have also considered that mean pooling of word embedding have outperformed complex models with LSTMs. Therefore they have explored the aspect of achieving higher accuracy with less complex models. They have extended the CBOW approach in Wor2Vec (Mikolov et al., 2013a) while introducing dynamic window and n-grams.Their main contribution is a model less complex, efficient and scalable and at the same time higher in performance.

2.3 Encoder-Decoder Models

The work of (Luong et al., 2015) elaborates the usage of Encoder-Decoder architecture based on RNNs for Machine Translation tasks. They have incorporated an attention mechanism on the output sequence of the Encoder to iteratively decode translated text for the given input. In another study (Datta et al., 2020), authors have used Recurrent Neural Networks (RNNs) in an encoder-decoder structure for neural machine translation. They have conducted the study for translating English into French Language.

2.4 STS Dataset

Semantic Textual Similarity (STS) is a measurement used to assess the closeness of two text content with respect to their semantic meaning. An evaluation toolkit for universal sentence representation is defined which makes use of STS dataset (Conneau and Kiela, 2018). STS Benchmark dataset (Cer et al., 2017) consists of pairs of sentences and the corresponding STS Scores manually annotated for each pair of sentences. STS score is a value between 0 and 5 where the perfect semantic similarity between two sentences is represented by the score of 5. Scores close to 5 represents the sentence pairs that are somewhat producing the same meaning while scores close to 0 represents irrelevant sentence pairs.

This dataset is used for training state-of-the-art sentence embeddings by (Reimers and Gurevych, 2019) through supervised learning approaches. They have shown that sentence embeddings trained using supervised learning perform significantly better in semantic text comparison tasks compared to sentence embeddings trained using unsupervised or self-supervised methods. The supervised training approach used by (Reimers and Gurevych, 2019) optimizes the model based on the difference between the cosine similarity of a sentence pair and the normalized STS score (within the range 0 to 1). The goal of this optimization is to move the vectors representing semantically similar sentences close in the high-dimensional vector space. Moreover, correlation results generated by evaluating models for STS Benchmark dataset is used in comparing the performance of sentence embeddings in general (Reimers and Gurevych, 2019; Huang et al., 2021).

3 Methodology

In this section, the data extraction process, preprocessing steps, word embedding training and sentence embedding training phases are discussed.

3.1 Dataset

The dataset used for the training of the embedding was extracted from the United States Supreme Court Case Law records extracted from FindLaw website[1]. The case law documents were chosen from Criminal Cases ranging from the year 2000 to 2010.

3.2 Pre-processing

Initially, the text files containing extracted court cases were processed to filter the body of the texts by removing title and footnotes sections in the documents. Stanford NLP python library: Stanza (Qi et al., 2020) is used to split sentences from the case texts. After observing some anomalies in split sentences, following pre-processing steps are applied to case paragraphs.

- Replaced abbreviations specific to legal domain with their long form

  Fed.R.Crim.P. – Federal Rule of Criminal Procedure

  Fed.R.Evid. – Federal Rule of Evidence

  Fed.R.Civ.P. - Federal Rule of Civil Procedure

- Removed non-ascii characters

- Removed content within rounded brackets if there are more than 2 words

  contained references and citations for legal documents

  no semantic meaning with respect to containing sentence

Following text pre-processing methods are applied to case sentences to make the text compatible for tokenization.

- Removed square brackets around letters and words

  Ex: [T]he, [petitioner], refer[s]

  Reason: caused due to styles used in web pages

- Removed numbering from the start of topic sentences

  Ex: I., A., II., 1.

[1] https://caselaw.findlaw.com/

- Replaced citations of previous cases with [CITE] keyword

  Ex: Pennsylvania v. Muniz, 496 U.S. 582, 601, 110 S.Ct. 2638

  Reason: reduce the distortion caused by citations for the semantic meaning of the sentence

- Removed sentences with more than 25% of [CITE] keyword with respect to all words

- Replaced continuous dashes, commas, white spaces with single entities

3.3 Word Embeddings for Legal Domain

Text corpus of 10,000 cases (extracted in section 3.1) containing more than 3 million words is used to train word embeddings. 300-dimensional Vectors are trained for 54059 unique words which appear more than 2 times within the corpus. Word2Vec (Mikolov et al., 2013a) and FastText (Mikolov et al., 2018) models are trained using Gensim library [2] and GloVe (Pennington et al., 2014) model is trained using glove_python [3] library. Window of 5 tokens before and after a token is used to specify the context when training the models.

3.4 Auto-Encoder Model

Due to the availability of a large in-domain text dataset, we searched for an unsupervised approach for learning sentence embeddings for the legal domain. We came up with the Auto-Encoder architecture, inspired by the application of Encoder-Decoder architecture in Neural Machine Translation systems (Datta et al., 2020; Luong et al., 2015). The objective of the Auto-Encoder is to reconstruct the original sentence token-by-token in an iterative manner using the state from previous tokens of the sentence and the vector representation for the whole sentence generated by the Encoder.

The workflow of the Auto-Encoder for a sentence containing m tokens at the $(k\text{-}1)^{th}$ iteration of the decoder is displayed in Figure 1. Upper section of the diagram represents the Encoder and lower section, the Decoder. The Embedding layers used in both Encoder and Decoder share same embedding matrix populated with pre-trained word embeddings.

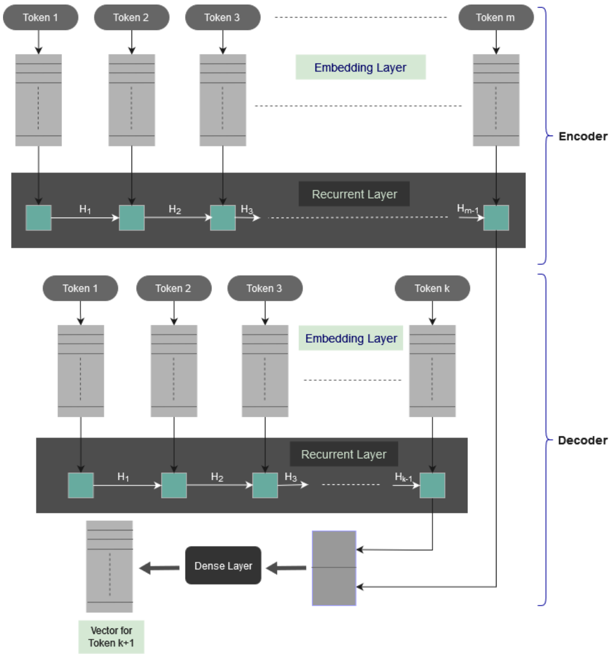


Figure 1: Auto-Encoder Architecture

Encoder takes in a sentence as a sequence of tokens and outputs a vector representation for the sentence. According to Figure 1, Embedding layer outputs a sequence of m vectors which is then passed on to a Recurrent layer with a specified number of units. The final state vector of the Recurrent layer is considered as the sentence embedding which is passed on to the decoder.

Each sentence is padded from the beginning with a *[START]* token and a *[END]* token to mark the beginning and the end of a sentence. Decoder iteratively predicts the next token starting from the *[START]* token at the first iteration to predict the token after the *[START]* token. Figure 1 depicts the $(k\text{-}1)^{th}$ iteration of the decoder, where the vector for $(k+1)^{th}$ token is predicted. Decoder takes in the k tokens preceding the $(k+1)^{th}$ token and passes them to the Embedding layer which outputs a sequence of k vectors. These vectors are passes on to a Recurrent layer where the final state vector is concatenated with the sentence vector provided by the Encoder. These concatenated output is passed on to a Dense layer which outputs a vector with the same dimension of the pre-trained embeddings.

Training loss is calculated at each decoding iteration, using the mean squared error between the predicted token vector and the actual vector ob-

[2] https://radimrehurek.com/gensim/

[3] https://github.com/maciejkula/glove-python/

tained from pre-trained word embeddings. Cosine similarity is used as the accuracy metric to evaluate the similarity between predicted and pre-trained word vectors.

The ability to predict the next token at each decoding step is based on the semantic meaning captured by the Encoder for the complete sentence and the state captured by the Decoder's Recurrent layer about the tokens preceding the to-be-predicted token of the sequence.

3.5 STS Dataset for Legal Domain

Understanding the need of a labeled dataset for legal domain to train and evaluate sentence embeddings. An STS dataset was prepared using sentences taken from US Supreme Court criminal cases and combining legal specific sentence pairs taken from STS Benchmark dataset with the assistance of a legal professional. Sample sentence pairs taken from the prepared dataset is displayed in Table 1.

Table 1: STS Legal Dataset - Samples

Sentence 1	**Sentence 2**	**Score**
Petitioner explained that his actions were taken in self-defense.	During the court proceedings, plaintiff argued he was only trying to save himself.	4.25
He did not present any evidence.	There were no evidence to support him.	3.25
The memorandum argued that plaintiff was not a risk to public safety and that he had accepted responsibility for his crime.	Court hearing raised concerns about the public safety.	1.5

First pair of sentences in Table 1 has a high STS score because the semantic meaning is same despite some content before the second sentence. Second pair of sentences has somewhat lower score since the first sentence doesn't elaborate the need for the petitioner to present evidence. It could be about supporting himself or against the opponent party. Last sentence pair has a low score since they are irrelevant despite the mention of public safety.

3.6 Multi-task Model for learning Sentence Embeddings

Since we have prepared a large dataset of case sentences and obtained a labeled dataset of STS score annotated sentence pairs, we focused on training a model for multiple tasks. To make use of the large set of unlabeled sentences, we defined a task to determine whether a sentence is distorted or not. Legal STS dataset is used for the task of predicting the similarity between two sentences and evaluating against the STS score. We list down the two tasks that the model is trained for:

- Noise added sentence discrimination
- Semantic similarity between a sentence pair

Noise addition process for sentences is done using a random word replacement algorithm. First, a set of general english words is extracted from the case sentence dataset. This set of words does not contain any person names, organization names or punctuation marks. Total number of general words accounts for 17796.

This set of words is used to replace 20% of words within each sentence by picking randomly. With this word replacement, the semantic meaning of the sentence is distorted. An example is displayed in Table 2.

Table 2: Noise Addition for Sentences

Original Sentence	**Distorted Sentence**
Plaintiff argued that the district court decision was unreasonable.	Plaintiff **guilty** that the district court **an** was unreasonable.

50% of the sentence dataset is distorted by random replacement and the label 1 is assigned for each distorted sentence. Label 0 is assigned for each original sentence.

According to Fig. 2, The model is trained for sentence discrimination task and sentence pair similarity task at each training step. Model shares the same embedding layer and Recurrent Neural Network (RNN) layer for both tasks. Sentence Dataset provides a batch of sentences containing original and distorted sentences and from the Dense layer output, the probability of a sentence being either original or distorted is calculated. Discrimination loss is then calculated using

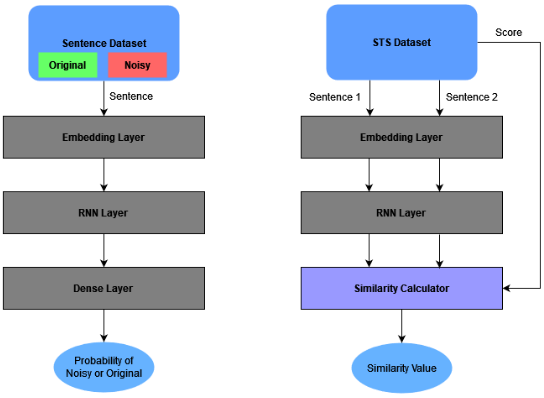


Figure 2: Multi-task model Architecture

this probability and the actual label. At the same training step, STS dataset provided a batch of sentence pairs and the similarity calculator produces the cosine similarity between the two vectors output by RNN layer. STS loss is calculated using the similarity value and the STS score provided by the dataset. Model weights are optimized using both Discrimination loss and STS loss.

This multi-task approach aims to train the model to capture the semantics of the sentences while preventing the model from over-fitting for STS Dataset which is relatively small compared to Sentence Dataset. Discrimination task force the model to identify distortions only by looking at the sentence vector provided by the RNN layer. STS task trains the model to move vectors of similar sentences closer in the vector space and irrelevant sentences further away. This approach is suitable for a setting where a large corpus of unlabeled data is available along with a small set of labeled data and the annotation cost is high to expand the labeled dataset.

4 Experiments and Results

In this section we discuss the variations that were done to identify most effective configurations. Several variations were experimented with the choice of word embeddings trained on general corpus and legal domain corpus. Also few variations were also tested with the auto encoder model to identify relatively better combination.

4.1 Pre-trained Word Embeddings

For our experiments, 3 types of word embeddings trained on general corpora and the legal corpus of 10,000 US Supreme Court cases, are used to initialize the Embedding Layer of the Auto-Encoder model. Pre-trained word embeddings on general corpora are obtained from online sources.

- Word2Vec
- Glove
- FastText

In the token distribution for 10,000 cases, 52% of the total sentences contain tokens within the range 20 - 40. We will be referring to the word embedding types trained on this legal corpus as *Word2Vec_Legal*, *GloVe_Legal* and *FastText_Legal* for the purpose of distinguishing them from word embeddings trained on general corpora.

4.2 Auto-Encoder Results

1000 US Supreme court cases consisting of 125719 sentences are used for the training and evaluation of the Auto-Encoder model. Experiments are done based on the word embedding type and Recurrent layer type. All the variations listed in Table 3 are trained for 20 epochs and measured the results as a controlled experiment to choose the relatively best variation for further training.

4.3 Multi-task Model Results

Multi-task Model of Noise Discrimination and STS tasks is trained and evaluated using different configurations of GRU layers. Accuracy and F1 scores are recorded for Noise Discrimination task and STS evaluation results are recorded using Pearson and Spearman Correlation between predicted similarity value and the STS score. Table 4 displays the results.

5 Conclusion and Future Work

Despite the availability of massive amount of text data, legal domain has inherent domain complexities and suffers from lack of annotated data. In this research we have conducted experiments with several variations to identify suitable sentence embedding models for the legal domain with this low resource settings. Self supervised approach is leveraged to overcome the lack of annotated data in the domain. This research serves as a preliminary step towards getting a proper numerical representation of a legal case. We intend to use the insights gained from this research to advance the sentence embeddings for more accurate results, with the use of relatively higher computational resources effectively.

Table 3: Model Variation Metrics

RNN Type	Units	Word Embedding	Train Cosine Sim.	Validation Cosine Sim.
GRU	512	Glove	0.2663	0.2578
		Word2Vec	0.2068	0.2033
		FastText	0.3323	0.3299
		Glove Legal	0.2776	0.2743
		Word2Vec Legal	0.2358	0.2310
		FastText Legal	0.2182	0.2153
Bi-GRU	512	Glove Legal	0.2550	0.2499
		Word2Vec Legal	0.2249	0.2180
		FastText Legal	0.2139	0.2097

Table 4: Multi-task Model Metrics

RNN Type	**Units**	**Accuracy**	**F1 (Original)**	**F1 (Noisy)**	**Pearson C.**	**Spearman C.**
GRU	512	92.21	91.94	92.46	46.81	48.09
GRU	768	93.70	93.71	93.70	57.62	51.34
LSTM	512	89.68	89.36	89.98	39.93	35.25
LSTM	768	92.73	92.52	92.92	34.35	28.76

We intend to make use of the derived sentence embedding models to legal domain specific tasks such as winning party prediction of legal cases.

References

[Arora et al.2017] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

[Cer et al.2017] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

[Cer et al.2018] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[Conneau and Kiela2018] Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

[Conneau et al.2017] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

[cor2020] 2020. Case law. https://www.law.cornell.edu/wex/case_law. Accessed: 2021-05-27.

[Datta et al.2020] Debajit Datta, Preetha Evangeline David, Dhruv Mittal, and Anukriti Jain. 2020. Neural machine translation using recurrent neural network. *International Journal of Engineering and Advanced Technology*, 9(4):1395–1400.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Huang et al.2021] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. *arXiv preprint arXiv:2104.01767*.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Luong et al.2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mikolov et al.2018] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[Pagliardini et al.2018] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Qi et al.2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

[Reimers and Gurevych2019] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wieting et al.2015] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

[Yang et al.2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

How Does Structure Affect Surface Rule Application?: Accent and Verbal Morphology in Japanese Deverbal (Noun) Compounds

Yu Tomita
Universität Leipzig, Germany
yt44kiry@studserv.uni-leipzig.de

Abstract

This paper investigates deverbal compounds in Japanese. Previous studies have discussed the classification of deverbal compounds, which can be classified according to their structure. Some former investigations discovered some generalizations on accents, *rendaku*, and direct verbalization. Here I propose a generalization on direct verbalization and deaccentuation, combining two previous generalizations. I also suggest that some deverbal compounds do not have accents due to a functional head, realized as a phonological floating feature on (de)accentuation, by utilizing the Distributed Morphology (DM) framework.

1 Introduction

Here I investigate morphological and phonological aspects of Japanese deverbal (noun) compounds[1]. These deverbal compounds consist of two components: non-head element (**dependent**) and verb-infinitive noun head (**base verb**), where the base verb stands for a vowel-final verb stem or consonant-final verb stem followed by the suffix *-i* (conjugation form)[2]. For example, (1a) and (1c) consist of the dependent and the consonant-final base verb *kir*, whereas (1b) consists of a dependent *azi* and a vowel-final base verb *tuke*.

(1) a. *hará* + *kirí* →
berry cutting
harakirí
'suicide with a sword'

b. *azi* + *tuké* → *ajituke*
taste adding 'seasoning'

c. *ne* + *kirí* → *ne**g**iri*
price cutting 'discounting'

d. *usu* + *kirí* →
thin cutting
*usu**g**iri*
'cutting into thin pieces'

e. *kuruma* + *tome* →
car stopping
*kuruma**d**ome*
'roadblock'

As shown in (1c–1e), NV compounds can undergo sequential voicing (***rendaku***). As well

[1] Here, I use the term **deverbal (noun) compounds** as referring to compounds of some word and a deverbal noun, setting aside the similar term **synthetic compounds**, which is generally understood to mean more specific compounds consisting of a noun and a (transitive) verb.

[2] In this paper, I employ the Japanese system of Romanization instead of phonetically rigorous transcription. See Yamaguchi (2014), e.g., for detail.

as other NN compounds, NV compounds generally obey (Motoori-)Lyman's law.

(2) **(Motoori-)Lyman's Law (Lyman 1894):** in a compound noun, if the second constituent contains voiced obstruent, it bleeds *rendaku*.

Semantically, these compounds denote various entities and eventualities. For instance, *harakiri* means an event of committing suicide, whereas *kurumadome* denotes some entity. These compounds behave as N(P)s but partially have verbal properties. Moving on to the verbal properties of deverbal compounds, some of them can be used like a verb with the light verb *suru*, as shown in (3a–3d).

(3) a. *Syogun-wa hara-kiri(-o)*
general-TOP berry-cutting(-ACC)
sita
did
'A *shogun* committed a suicide.'

b. *Yiding-wa banana-ni*
Y-TOP banana-DAT
azi-tuke(-o) suru
taste-adding(-ACC) do.PRES
'Yiding flavors banana.'

c. *Elin-wa kagu-o*
E-TOP furniture-ACC
ne-giri(-o) suru*
price-cutting(-ACC) do.PRES
'Elin haggles furniture.'

d. *Lukas-wa niku-o*
L-TOP meat-ACC
*sen-giri(-ni/*o)*
thousand-cutting(-DAT/ACC)
suru
do.PRES
'Lukas shreds meat.'

e. * *Mary-wa kurumadome(-o)*
M-TOP car-stopping(-ACC)
suru
do.PRES
Intended: 'Mary parks.'

However, other deverbal compounds which only denote entities or resultant of events (result nominals, Grimshaw 1990) cannot be used like a verb, as shown in (3e). Following Ito & Sugioka (2002), I will focus on deverbal compounds denoting eventualities.

I will analyze these deverbal compounds with Distributed Morphology (DM) framework (Halle & Marantz 1993), focusing on deverbal compounds' phonological and morphological behaviors. For example, some combination of a dependant and a base verb results in two compounds that are phonologically and semantically different. (4) exhibits the different compounds from the same dependent *zin* and base verb *tori* (Tatsumi 2016, 2021).

(4) *zin* + *tori* 'spot + take'

a. *zintóri* 'Tom Tiddler's ground'

b. *zin**d**ori* 'encamping'

It indicates that there is a morphological difference between these compounds.

The rest of the structure in this paper is as follows. First, I review previous work on Japanese deverbal compounds. Especially, I focus on Yamaguchi (2014) and Tatsumi (2016, 2021), which propose useful generalization of deverbal compounds. Then I will propose a generalization of deverbal compounds in accents and direct verbalization. I will integrate their ideas. Finally, I will suggest an analysis of deverbal compounds in the DM framework. Following Volpe (2005), I will classify the structure base verbs into two subcategories. Then I will analyze the deaccented deverbal compounds due to some functional head, which is realized as a phonological floating feature on (de)accentuation, by utilizing the DM framework.

2 Previous studies

The previous investigations on deverbal compounds have mostly focused on NV compounds and their phonological aspects, especially *rendaku* and accentuation. Some earlier papers discovered a phonological and morphological generalization of these compounds.

2.1 Argument and Adjunct Type

Sugioka (2002) and Ito & Sugioka (2002) classify deverbal compounds (containing bimoraic base verbs) into two types. One is called **(Direct) Argument type**, in which the dependent corresponds to the internal argument, as in (5).

(5)	*mádo*	+ *hukí*	→ *madóhuki*
	window	wiping	wiping window

The other is called **Adjunct type**, in which the first element does not correspond to the internal argument, as in (6).

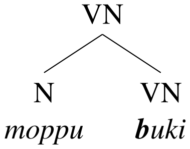
(6)	*moppu*	+ *hukí*	→
	mop	wiping	

*moppu**b**uki*
wiping with a mop

Adjunct Type

The dependent modifies the base verb as an adjunct in these deverbal compounds. Sugioka (2002) and Ito & Sugioka (2002) argue that this kind of deverbal compound generally undergoes *rendaku* and is deaccented, as shown in (6). They argued that these deverbal compounds contain verbal noun as shown in below.

(7)

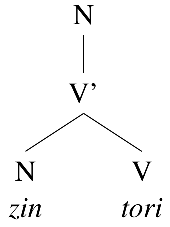


→ *moppu**bu**ki*

Argument Type

Sugioka (2002) and Ito & Sugioka (2002) argue that this kind of deverbal compound tends to undergo *rendaku* but have an accent, as shown in (5). They analyzed these deverbal compounds as the nominalized V' rather than compounds of dependents and nominalized base verbs.

(8)



→ *zintóri*

[3]

2.2 Corpus Studies

The analysis in Ito & Sugioka (2002) and Sugioka (2002) is supported by subsequent corpus studies, but they also obtained some results not contained in previous work.[4].

Complementary Distribution between Accents and *Rendaku*

Yamaguchi (2014) surveyed some corpora and found a complementary distribution be-

[3] However, as (1c) illustrates, there are Argument type deverbal compounds that undergo *rendaku*, as shown in (i).

(i) Argument type deverbal compounds which undergo *rendaku* (Tagawa 2010)

a. *misé* + *simai* → *misezímai* 'store + put away' → 'closing a store'
b. *úmi* + *hirakí* → *umibíraki* 'sea + open' → 'the beginning of a sea-bathing season'
c. *tikará* + *soe* → *tikarazoe* 'power + attach' → 'helping'
d. *kuti* + *tuké* → *kutizuke* 'mouth + put' → 'kissing'
e. *toogé* + *koe* → *toogegoe* 'peak + go over' → 'crossing over a peak'
f. *sina* + *kiré* → *sinagire* 'goods + go over' → 'being out of stock"

[4] Note that corpora used in these studies contain deverbal VV compounds, in which the dependent is also a deverbal noun. This may affect the tendency of *rendaku*, but it has something little in common with this paper.

tween *rendaku* and accentuation in deverbal compounds containing the bi-moraic base verb.

(9) **Yamaguchi's complementary distribution**
If a base verb in a deverbal compound is bi-moraic, then *rendaku* and accentuation exhibit complementary distribution.

This generalization states that Japanese deverbal compounds always carry an accent or undergo *rendaku* but not both.

Also, Yamaguchi (2014) discovered that some deverbal compounds with tri-moraic base verbs do not obey this distribution. That is, deverbal compounds with trimoraic base verbs strongly tend to have accents.

Tendency of *Rendaku*

Fukasawa (2020) carried out a corpus study and found the base verbs with particular tendencies to undergo *rendaku* in (external- and internal-) argument type environment, which is summarized in (10).

(10) a. −R verbs (tend to avoid *rendaku*):
kiri 'cutting', *tuki* 'attaching', *kakusi* 'hiding', *tataki* 'hitting', ...

b. +R verbs (tend to undergo *rendaku*):
tome 'stopping', *kaesi* 'returning, *kaki* 'writing', *kosi* 'passing', ...

According to her statistical result, VV compounds also tend to avoid *rendaku*. However, the avoidance of tendency in these compounds is weaker than in Argument type compounds.

2.3 DM Approaches

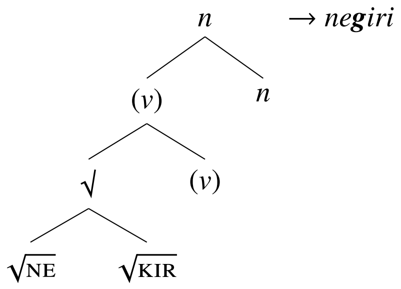
Moreover, some deverbal compounds can be used as a verb without the light verb *suru*. Instead, they can be directly verbalized in the sentence.

(11) a. *Hanako-wa syoohin-o*
H-TOP grocery-ACC
negiru
bargain.PRES
'Hanako beats down the price of groceries.'

b. *Taro-wa sakura-no sita-o*
T-TOP sakura-GEN bottom-ACC
zindoru
encamp.PRES
'Taro stakes out a spot under a cherry blossom.'

Unlike the examples in (3), in each sentence of (11) deverbal compounds are used as a verb directly. I will call these verbal use **verbal derivative**, to distinguish it with light verb construction. Tatsumi (2016, 2021) analyzed these deverbal compounds within phase-based Distributed Morphology (Arad 2003). In his analysis, all Adjunct type deverbal compounds and some Argument type deverbal compounds with *rendaku* contain a constituent that consists of at least two Roots without any intervening phase head, shown in (12).

(12)



Then he proposed a generalization on the verbal use of NV compounds.

(13) Tatsumi's (2016) Observation
If a noun-verb deverbal compound is used as a verb, it shows sequential voicing unless it violates Motoori-Lyman's Law[5].

[5] Tatsumi (2016, 2021) pointed out that there are some

This means that if a direct argument type compound can be verbalized without a verbalizer *suru*'do', it undergoes *rendaku*.

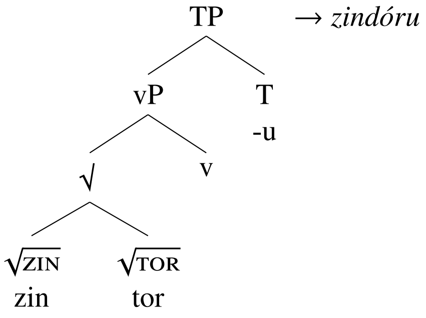
(14) a. *ne* + *kirí|kíru* →
price cutting|cut.PRES
*ne**g**iri|ne**g**íru*
'bargaining'|'bargain'

b. *ziń* + *torí|tóru* →
spot taking|take.PRES
*zin**d**ori|zin**d**óru*
'encamping'|'encamp'

c. *awá* + *tatí|tátu* →
bubble standing|stand.PRES
*awa**d**ati|awa**d**átu*
'lathering'|'bubble'

This structure allows direct verbalization.

(15)



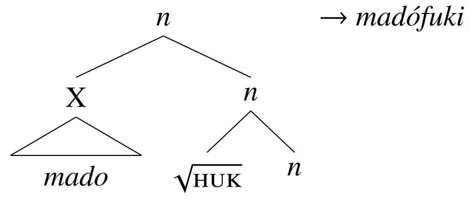
If a noun-verb deverbal compond does not undergo *rendaku*, then the deverbal compound cannot be used as a verb without *suru*.

(16) a. * *harakiru* (cf. *harakirí(-o)-suru*)

b. * *zintoru* (cf. *zintóri(-o)-suru*)

He also suggests that other Argument type compounds without *rendaku* can have a different structure, as shown in (17). This compound does not undergo *rendaku* since a projection of the phase head *n* intervenes between the dependent and the base verb.

(17)



He argues that this structure allows compounds neither to undergo *rendaku* nor to have verbal use. Hasegawa & Oseki (2020) also proposed a very similar analysis, arguing that a two-Root constituent does not bare an accent[6].

2.4 Interim summary

There are two findings on deverbal compounds.

(9) **Yamaguchi's complementary distribution**
If a base verb in a deverbal compound is bi-moraic, then *rendaku* and accentuation exhibit complementary distribution.

(but few) exceptions for this observation.

a. *kosi* + *kake* → *kosikáke*
hip hanging 'stepping stone'

For instance, the above example does not undergo *rendaku* but allows direct verbalization *kosikakéru*. In this paper, however, it is out of focus.

[6] These analyses seem to assume the intuition that closely connected constituents feed (sequential) voicing. This implies that if the two Roots are sisters, then the second one undergoes (sequential) voicing if morphophonological restrictions such as Motoori-Lyman's law are respected. Crucially, there are counterexamples against this conclusion, where only a two-Roots constituent undergoes *rendaku*. They evidently contain a non-Root dependent.

(i) *mi-kake* + *taosi*→ *mikake**d**aosi* 'appearance + bringing down; not as good as it looks'

(ii) *kaw-ar-i* + *hae*→ *kawari**b**ae* 'replacement + good looking; improvement'

As for (i), *mikake* consists of at least two stems; *mi* and *kake*. If we assume Tatsumi's hypothesis and analyze *mikake* as a single Root, then each *mi* and *kake* must be a Root, and compounds of them should undergo *rendaku*. If *mikake* is not a single Root, then it might be predicted that *taosu* does not undergo *rendaku*. However, both possible analyses fail to predict the form *mikakedaosi*. Similarly, since *kaw-ar-i* contains an intransitive marker *-ar*, the dependent in (ii) is not a Root.

(13) **Tatsumi's observation**
If a noun-verb compound allows direct verbalization and its base verb is *rendaku*-capable, the compound undergoes *rendaku*.

These generalizations seem to complement each other, but you can find several exceptions against both of them: deverbal compounds shown in (18). They do not undergo *rendaku* or have an accent but allow direct verbalization.

(18) NV compounds allowing direct verbalization

a. *azi* + *tuké|tukéru* →
taste attaching|attach.PRES
azituke|azitukéru
'seasoning'|'flavor'

b. *maru* + *tuké|tukéru* →
circle attaching|attach.PRES
marutuke|marutukéru
'grading'|'mark as correct'

c. *húu* + *kirí|kíru* →
seal cutting|cut.PRES
huukiri|huukíru
'{releasing|release} (a film)'

d. *áku* + *taré|taréru* →
evil dropping|drop.PRES
akutare|akutaréru
'verbal abuse'|'behave badly'

e. *kotó* + *kaki|káku* →
thing lacking|lack.PRES
kotokaki|kotokáku
'shortage'|'lack'

In the next section, I will propose an integrated generalization which can cover these examples.

3 Integrated generalization

I do not offer any new observations or analyses of *rendaku* but discuss a relationship between direct verbalization and accents. Yamaguchi's complementary distribution entails that if a deverbal compound undergoes *rendaku*, it also undergoes deaccentuation. Then I propose an integrated generalization of the statements from Yamaguchi's distribution and Tatsumi's observation.

(19) **Yamaguchi-Tatsumi's generalization (to be revised)**
If a noun-verb compound allows direct verbalization and its base verb is *rendaku*-capable and bimoraic, then the compound (denoting eventualities) has no accent.

The examples in (14), which obey Tatsumi's original observation, also follow the proposed generalization. Besides, some of the counterexamples against Tatsumi's observation are subject to this generalization. The examples in (18) contain *rendaku*-capable base verbs, and NV compounds in (18) allow direct verbalization.Tatsumi's observation does not predict (18) since they do not undergo *rendaku*. However, (18) obey the generalization (19) since they are deaccented.

The generalization is also applicable to many compounds consisting of nominal dependents and *rendaku*-incapable base verbs, which allow direct verbalization.

(20) X = N cases

a. *na* + *nori|noru* →
name riding/ride.PRES
nanori|nanóru
'{giving|give} one's name'

b. *katá(ho)* + *yori|yoru*
one side approaching/approach
→ *katayori|katayóru*
'deviation'|'lean over'

Moreover, the proposed generalization covers a lot of XV compounds (X = V, Adj), which allow direct verbalization. Especially, according to Fukasawa (2020), VV compounds tend to avoid *rendaku*, although their behavior is similar to adjunct type deverbal compounds. This tendency is weaker than Argument type VCs but not covered by Tatsumi's original observation. It is, however, compatible with my proposal: Yamaguchi-Tatsumi's generalization.

(21) X = V cases

a. *kiki* + *torí*|*tóru* →
listening taking|take.PRES
kikitori|*kikitóru*
'{hearing|hear} what others say'

b. *uti* + *kirí*|*kíru* →
hitting cutting|cut
utikiri|*utikíru*
'finish'

c. *mi* + *hari*|*haru* →
seeing spreading|spread.PRES
mihari|*miharu*
'watch'

(22) X = Adj cases

a. *chiká(-i)* +
near
yori|*yoru* →
approaching|approach.PRES
chikayori|*chikayóru*
'getting closer'|'go closer'

b. *taká(-i)* + *nari*|*náru* →
high ringing|ring.PRES
takanari|*takanáru*
'fast beating'|'beat loudly'

Therefore, (19) will be revised in the following way.

(23) **Yamaguchi-Tatsumi's generalization (final)** If an XV compound allows direct verbalization and its base verb is bimoraic, then the compound (denoting eventualities) has no accent.

Note that deverbal compounds which do not undergo *rendaku* or have accents, shown in (24), do not obey Yamaguchi's (2014) complementary distribution. However, they are no longer counterexamples against the proposed generalization (23).

(24) Deaccented deverbal compounds without *rendaku*

a. *ika* + *turi*→ *ikaturi* 'squid + fishing; fishing squids'

b. *kuti* + *kirí*→ *kutikiri* 'mouth + cutting; opening'

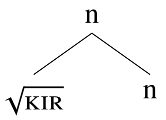
In summary, the lastly proposed generalization covers these XV compounds regardless *rendaku*-capability:

- Argument type NV (tend to avoid *rendaku*)
- Adjunct type NV (tend to undergo *rendaku*)
- VV (tend to avoid *rendaku*)
- AdjV

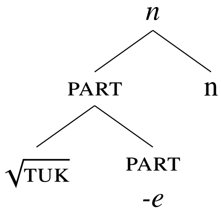
4 Towards a DM analysis

Following Volpe's (2005) analysis of Japanese verbs, I assume that a base verb consists of several segments, including Roots, categorizer heads such as *n* and *v* and Affixal Particle head PART (den Dikken 1995), introducing Root-derived and verb-derived distinction. All morphological segments of each base verb shown in (1a) and (1b) are in (25a) and (25b), respectively. In (25b), *-e* is a transitive marker, while the suffix *-i* in (25a) is analyzed as phonological epenthesis with consonant-final Roots (Poser 1984).

(25) a. Root-derived: *kir*√-*i*(-øv)-øn

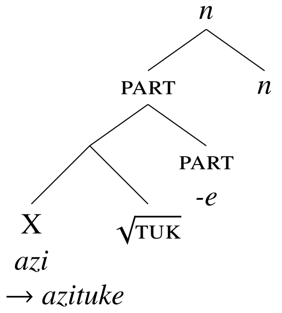


b. Verb-derived: *tuk*√-*e*v-øn



Then, the example (1b) *azituke* 'seasoning' can be analyzed as shown in (26).

(26)



Going back to (10), Fukasawa's findings on the tendency of *rendaku* in base verbs, −Rendaku verbs such as *kiri*, *tuki*, and *tataki* are possibly Root-derived since they do not have any overt transitivity morphology (*kiri* and *tataki* cannot alter intransitive forms; *tuki* has a transitive alternation *tuke* and *tukasi*), while most +Rendaku verbs such as *tome* and *kaesi* overtly contain transitive markers, indicating that they are verb-derived[7].

Following Trommer (2019), I will argue that some floating feature H subtracts an accent in deverbal compounds. I assume that H is an exponent of some functional category. Based on the findings in Yamaguchi (2014), it seems straightforward to think that *rendaku* and deaccentuation have equivalent initial status, and that PART can be realized as such an exponent involving lexical insertion of some (overt) transitivity morpheme.

(27) Possible analysis (to be revoked):
PART ↔ H + /-*e*/

However, the verbal derivatives shown in all previous examples except (21c) have accents, even though deverbal compounds are deaccented. Therefore, the above approach does not work. Instead, I propose that the *n* is realized as a floating featural element H that eliminates accents in a certain environment.

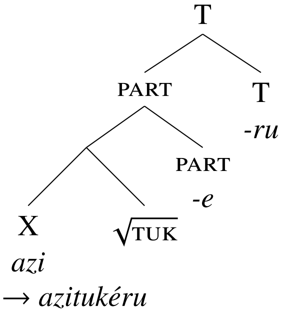
(28) Proposal: contextual allomorphy

a. H ↔ *n* / PART

b. ø ↔ *n* / elsewhere

This context-sensitive exponent in (28) allows deverbal compounds to keep their original accents after direct verbalization since verbal derivatives do not contain *n*[8].

(29)



[7] Note that though *kakusi* also contains a transitive marker, it exceptionally avoids *rendaku* in modern Japanese even if it appears in Adjunct type environment, as shown in (i).

(i) *hita(sura)* + *kakusí* →
solely hiding
hitakakusi
'hiding (a secret) at all costs'

In old Japanese, however, *kakusi* was allowed to undergo *rendaku*.

(ii) *omo* + *kakusi|kakusu* →
face hiding|hide
omogakusi|omogakusu
'{hiding|hide} one's face'

Tatsumi's (2016) observation correctly predicts the *rendaku* in (ii) since it allowed direct verbalization.

[8] Oseki (2017) provides a more sophisticated analysis of Japanese verbal morphology in the DM framework. I leave the development of the proposed research incorporating his idea as future work.

Note that as mentioned in Volpe (2005), bi-moraic verbs are hard to use solely as a nominal with verb-related meaning. Even though they are used in conjugation form and they still have an accent, deverbal compounds are different from them.

5 Limitations

Although the coverage of Yamaguchi-Tatsumi's generalization is restricted to deverbal compounds with bi-moraic base verbs, some exceptions do not obey the proposed generalization.

(30) Exceptional accented deverbal compounds allowing direct verbalization:

a. *yokú* + *hari|haru* →
greed spreading|spread.PRES
yokubári/yokubáru
'grasping'|'be greedy'

b. *kosi* + *kaké|kakéru* →
hip hanging|hang.PRES
kosikáke|kosikakéru
'seating'|'take (chair)'

However, one of the two previous generalizations still predicts them. Tatsumi's observation correctly predicts (30a), and (30b) obeys Yamaguchi's distribution. Logically speaking, you cannot find a counterexample that violates all generalizations since when an accented deverbal compound undergoes *rendaku*, it cannot be a counterexample of Tatsumi's observation.

6 Conclusion

In this paper, I investigated Japanese deverbal compounds and proposed Yamaguchi-Tatsumi's generalization, arguing that deverbal compounds which allow direct verbalization does not have any accent. This property can be explained in the DM framework by utilizing contextual allomorphy.

The important problem is that most Japanese deverbal compounds do not allow direct verbalization, unlike English synthetic compounds. Tatsumi (2016, 2021) argued that most of the argument type deverbal compounds contain *n* as a unique functional head, indicating that it is only used as a noun. However, many other compounds which undergo *rendaku* and are deaccented do not allow direct verbalization.

This paper and previous articles cited in this paper mainly focused on NV compounds. As you have seen, however, the proposed generalization (23) is applicable to lots of deverbal XV compounds. Further investigation should be carried out on (de)accentuation in deverbal noun compounds, including VV and AdjV compounds. Especially, VV compounds have been widely investigated, but there is little connection between research on NV compounds and VV compounds.

Acknowledgements

I would like to thank the member of the Institute of Linguistics at Leipzig University for their helpful comments and discussions. I would like to thank Hitomi Hirayama and Kenji Oda for their personal meeting. Finally, I would like to thank two reviewers in PACLIC 2022 for their comments. All remaining errors are my own.

References

Arad, Maya. 2003. Locality Constraints on the Interpretation of Roots: The Case of Hebrew Denominal VERBS. *Natural Language & Linguistic Theory* 21(4). 737–778. https://doi.org/10.1023/A:1025533719905.

den Dikken, Marcel. 1995. *Particles: On the Syntax of Verb-Particle, Triadic, and Causative Constructions*. Oxford University Press.

Fukasawa, Michiko. 2020. Rendaku in the Syntax-Phonology Interface: A Corpus-Study on Deverbal Noun Compounds. In Michael Barrie (ed.), *Japanese/Korean Linguistics*, vol. 27. CSLI Publications.

Grimshaw, Jane. 1990. *Argument structure* (Linguistic Inquiry Monographs 18). Cambridge, Mass: MIT Press.

Halle, Moris & Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In *The view from Building 20: Essays in Linguistics in Honor of Sylvian Bromberger*, 111–176. Cambridge, Massachusetts: MIT Press.

Hasegawa, Takuya & Yohei Oseki. 2020. Bunsanketairon to Nihongo no Doshi Yurai Fukugo Go [in Japanese]. In *Handbook of the 161th meeting of the Linguistics Society of Japan*, 299–305.

Ito, Takane & Yoko Sugioka. 2002. *Structure of words and word formation [in Japanese]*. Kenkyusha.

Lyman, Benjamin Smith. 1894. Change from surd to sonant in Japanese compounds. *Oriental Studies*. Oriental Club of Philadelphia. 160–176.

Oseki, Yohei. 2017. *Voice morphology in Japanese argument structures*. New York University Qualifying Paper.

Poser, William John. 1984. *The phonetics and phonology of tone and intonation in Japanese*. Massachusetts Institute of Technology dissertation.

Sugioka, Yoko. 2002. Incorporation vs. modification in deverbal compounds. In *Japanese/Korean Linguistics*, vol. 10, 495–508. CSLI Publications.

Tagawa, Takumi. 2010. A descriptive note on deverbal compounds in Japanese: with emphasis on their phonological behavior [in Japanese]. *Bulletin of Tsukuba Gakuin University* 5. 157–163.

Tatsumi, Yuta. 2016. Deverbal Compounds in Japanese: A Distributed Morphology Approach. In *Proceedings of Formal Approaches to Japanese Linguistics*, vol. 8, 177–187. Mie.

Tatsumi, Yuta. 2021. Structural restrictions on sequential voicing in Japanese N-V compounds. The 29th Japanese/Korean Linguistics Conference, Nagoya, Japan.

Trommer, Jochen. 2019. A Concatenative Account of Japanese Subtractive Accent. Universität Leipzig.

Volpe, Mark Joseph. 2005. *Japanese Morphology and its theoretical consequences: Derivational morphology in Distributed Morphology*. Stony Brook University dissertation.

Yamaguchi, Kyoko. 2014. *Accentuation and Rendaku in Japanese Deverbal Compounds: A Comparison with Noun Compounds*. University of Tokyo dissertation.

Collection Methods and Data Characteristics of the PagkataoKo Dataset

Edward Tighe, Luigi Acorda, Alexander Agno II, Jesah Gano,
Timothy Go, Gabriel Santiago, Claude Sedillo
De La Salle University, Manila, Philippines
{edward.tighe, luigi_acorda, alexander_agno, jesah_gano,
timothy_go, gabriel_santiago, claude_sedillo}@dlsu.edu.ph

Abstract

We present the PagkataoKo Dataset, a new dataset for Filipino Automatic Personality Recognition (APR) containing demographic, personality trait, and social media data sourced from 3,128 Filipino Instagram and/or Twitter users. As APR is focused on processing an individual's observable actions, we improve upon the previous Filipino APR dataset by collecting multimodal data, as well as similar data expressed in different environments. In our paper, we describe our collection methodology and detail the general characteristics of the dataset. We also report the language characteristics of the posts and highlight the presence of multiple languages (e.g. English, Tagalog) and code-switching – two aspects that make Filipino APR difficult. Lastly, we discuss how our dataset provides future work with multiple options to explore in order to navigate around the complexities found in Filipinos' language usage.

1 Introduction

Automatic Personality Recognition (APR) is a computing task that focuses on personality traits (i.e. individual differences) and their externalizations (Vinciarelli and Mohammadi, 2014). The task is rooted in the idea that traits influence a person's interactions in different environments (Larsen and Buss, 2008); hence, a person's observable actions contain traces of their traits, such as how it is believed that important personality characteristics and individual differences are encoded into one's language (i.e. the lexical hypothesis) (Goldberg, 1981; Tausczik and Pennebaker, 2010). As a computing task, APR, therefore, involves the collection and processing of these observable actions and analysis (e.g. descriptive, causal, predictive) of any traces of personality left behind.

APR has received much attention over the past two decades – leading to the exploration of many different types of mediums (e.g. text, image, audio) in which personality may have been expressed upon. Early work in APR mostly focused on studying language usage – with data coming from conversation recordings (Mehl et al., 2006; Mairesse et al., 2007), emails (Gill and Oberlander, 2002), and essays (Pennebaker and King, 1999; Argamon et al., 2005; Mairesse et al., 2007). These early studies did not produce high-performing predictive models nor did they have a high volume of data to validate results with but were at least able to show that indicators of personality can be found in one's language. APR studies then branched out and explored other sources of observable actions with a vast number of studies gravitating towards social media platforms, such as (but not limited to) Facebook (Golbeck et al., 2011b; Gosling et al., 2011; Wald et al., 2012; Markovikj et al., 2013; Schwartz et al., 2013; Kosinski et al., 2014; Park et al., 2015; Segalin et al., 2017), Twitter (Golbeck et al., 2011a; Quercia et al., 2011; Rangel Pardo et al., 2015; Liu et al., 2016; Skowron et al., 2016; Ong et al., 2017; Samani et al., 2018), Instagram (Ferwerda et al., 2015; Skowron et al., 2016; Lay and Ferwerda, 2018), Sina Weibo (Gao et al., 2013; Guntuku et al., 2015), Flickr (Samani et al., 2018), and general blogs (Nowson and Oberlander, 2006; Nowson and Oberlander, 2007; Gill et al., 2009; Yarkoni, 2010). Social media platforms have since become a perfect source of data for APR given the many different ways a person might interact within the online environment.

While more recent APR studies have gravitated towards exploring neural network based methods (Mehta et al., 2019), an area of opportunity within APR that lacks attention is in studying social media data from Filipinos. Kemp (2021) noted that individuals from the Philippines spent the most time on social media – clocking in a little over four hours a day on social media versus the global average of roughly 2.5 hours. This high usage is an indicator that there is a high volume of observable actions that can be collected from online Filipino users. However, despite this upside, social media data from Filipinos can generally be considered hard to deal with when coming from a natural language processing perspective. Filipinos are known to speak multiple languages (e.g. English, Filipino, Cebuano, and a number of other Philippine languages) and code-switch between these languages (Caparas and Gustilo, 2017; Abastillas, 2018; Tighe and Cheng, 2018). A corpus containing these language characteristics – coupled with informal language usage usually found in social media and the low number of language resources available for Philippine language processing – presents quite a challenge when looking to extract useful linguistic information related to personality.

Currently, only the dataset of Tighe and Cheng (2018) is suitable for Filipino APR. Tighe and Cheng (2018) produced a dataset containing text data from $610,448$ tweets of 250 Filipino Twitter users and were able to show that there were indeed some traces of Conscientiousness and Extraversion from term frequency inverse document frequency (TFIDF) values. However, a follow-up study by Tighe et al. (2020) showed that tuned multilayer perceptron (MLP) models trained on word embedding data (pre-trained and trained-over) did not learn at all and performed poorly when compared against MLPs using TFIDF. One reason for the poor performance can be attributed to the limited size of the data given their train-test split led to an even smaller amount of data for learning. It should be taken into consideration that the methods of Tighe et al. (2020) performed a limited analysis of the usage of word embeddings – implying that more detailed studies need to be crafted to gauge the usefulness of embedding-based approaches on Filipino text data. Nevertheless, the dataset's small size poses a limitation when applying certain computing methods (e.g. neural network approaches). In addition to the small data size, the dataset only contains text data from one platform. Related literature has shown that image data can also contain personality traces and aid in modeling personality (Liu et al., 2016; Segalin et al., 2017; Lay and Ferwerda, 2018). Also, a fusion of data – whether from different types of data (e.g. image + text + account) and/or different sources of data (e.g. Twitter + Instagram) – has produced better personality models against models using a single modality or sources (Skowron et al., 2016; Samani et al., 2018). The potential benefits of exploring different data modalities and sources is an aspect that the current Filipino APR dataset cannot provide to any future studies.

To address the need for a larger and more flexible data resource for Filipino APR, we created the PagkataoKo Dataset. The dataset contains personality, demographic, text, image, and account data from $3,128$ Filipino Instagram and/or Twitter users. Participants were administered the Big Five Inventory (BFI-44) to assess their trait scores and were given the choice to share access to one or both of their social media accounts' data. The novelty of our dataset lies in that the data is multimodal – capturing more observable actions than the previous dataset – and is sourced from two different platforms – capturing actions expressed in two different environments. In our paper, we discuss the methodology for collecting the data and detail general data characteristics, as well as temporal and language characteristics of the collected posts. We also discuss design considerations based on the characteristics of the data.

2 Collection Methodology

We extend the methodology of Tighe and Cheng (2018) by also collecting image and account-related data, aside from text data. We also gave participants an option to share access to multiple social media accounts, instead of just one. We selected Twitter and Instagram as the sources of observable actions because of how each platform encourages different behavior – with Twitter being micro-blogging oriented and Instagram being media-sharing oriented.

2.1 Personality Trait Representation

To assess trait scores, we used the Big Five Inventory (BFI-44), a 44-item self-reported questionnaire that measures the five dimensions of the Big Five (John et al., 1991; John et al., 2008). These five dimensions – sometimes collectively referred to as OCEAN – include Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each dimension or broad trait is associated to about 8 to 10 Likert scale questions ranging from 1 (strongly disagree) to 5 (strongly agree). Computation of a trait's score involves reversing specific item scores and then computing the average score of a trait's associated items – resulting in a value that ranges from 1.0 to 5.0. We selected the BFI for our study mainly because numerous other APR studies utilized a questionnaire that measured the Big Five. By standardizing the representation of personality to the Big Five, any results from our data would be easier to compare with other studies.

2.2 Collection Tool

We created a web application that facilitated the collection of demographic data, social media data, and personality trait scores of our participants. The application first presented individuals with the initial collection formalities (e.g. consent form, directions). After agreeing, participants would be asked to grant our application permission to read their social media data from either or both Twitter and/or Instagram. The application would then collect social media data by interfacing with each platform's respective APIs. Participants were then presented a demographic questionnaire followed by a personality test. The demographic questions were presented to acquire the information we need to describe and eventually filter the participants. Only individuals 18 years old or above were permitted to proceed with the collection. As for the personality test, we implemented an online version of the BFI-44. The application collected answers to each of the 44 items, as well as the computed Big Five scores. Participants were then shown their personality scores compared against the trait scores presented in Tighe and Cheng (2018).

For Twitter, our application would utilize the Twitter API v1 to collect an individual's most recent $3,200$ tweets – a limitation of the API. The application then discarded all retweets as we were only concerned with tweets that were written by the user; however, quoted tweets were retained since these were written by the user. Aside from tweet-level data, the application also collected links to each participant's profile picture, as well as other account-related data (e.g. # of followers / following). As for Instagram, our application would utilize the Instagram Legacy API to collect as many of a user's posts as the API would return. For each post, our application collected both the caption and a link to the post's image. In the case that a post had multiple photos, only the main photo was collected. As for posts that contained videos, our application discarded them as video data was not initially factored into the research design. In addition to the post data, our application collected a link to the profile picture and other account-related data to their Instagram account.

Links to images were collected instead of downloading the image itself to lessen the strain on the web application especially as Instagram allowed for the collection of all a user's posts. Our initial plan was to wait until the collection was over before downloading all images; however, this was a costly decision as we did not realize that the URLs from both platforms changed over time. The changing of an image's URL could be due to a user uploading a new profile picture after their participation or the platform periodically refreshing links. Because of this, we lost access to image data from around 350 Instagram and 804 Twitter users. Once we discovered this issue, we opted to continue retrieving image links but we would download images at the end of each day until the end of the collection.

2.3 Time Frame and Sampling Methods

Our collection started on the 1st week of Jun. 2019 and ended on the 2nd week of Feb. 2020. We utilized a mixed sampling approach centered on volunteer sampling. We approached individuals with idea that participating and disseminating word of the research would generally be beneficial to the research project. While not always highlighted in the invites/advertisements, we would compute and present the results of the personality test to each participant who finished the entire collection – a factor that proved useful in encouraging individuals to share the collection with their own networks as we did not offer incentives.

From the start of the collection until the 1st week of Oct. 2019, we performed convenience sampling by posting information about the recruitment within our immediate networks. Postings were made on different online platforms and by reaching out to individuals in person. We also achieved a minor snowball effect as our networks helped propagate the recruitment to their own respective networks. These initial efforts resulted in the collection of data from 362 participants. After which, we started online advertisement campaigns to reach out to Filipinos outside of our immediate network. We create ad campaigns that targeted individuals from the Philippines and had these ads run on Facebook, Instagram, and Twitter. We ran ad campaigns intermittently across the 2nd to 4th week of October 2019. This resulted in an additional $1,393$ individuals. The initial success of recruiting participants through ads led us to run additional ads throughout the 1st week of Dec. 2019 and the 2nd week of Feb. 2020. By the end of our collection, we were able to recruit and collect data from $3,186$ participants.

2.4 Participant Filtering

Participants were included in the final dataset if they signified that their nationality was Filipino. We also included individuals who were mixed Filipinos or individuals who stated they were Filipino and had one or more additional nationalities. After filtering, we were left with a total of $3,128$ individuals. We discarded data from those who did not qualify based on this filter.

2.5 Ethical Clearance

Our methods were reviewed and given clearance by the Research Ethics Office of De La Salle University, Philippines. Individuals voluntarily gave electronic consent to participate in our study and their social media data was collected in accordance with the developer policies of Twitter and Instagram.

3 The PagkataoKo Dataset

The PagkataoKo[1] Dataset is composed of demographics, personality trait scores, user-generated data (e.g. post, tweet, profile pictures), and other account-related metadata from $3,128$ Filipino Instagram and/or Twitter users.

3.1 Subgroups of Participants

We organize participants into four subsets based on the social media account(s) they provided as follows:

- **I** – All participants with Instagram accounts,
- **T** – All participants with Twitter accounts,
- **I**∪**T** – All participants (i.e. the union between **I** and **T** or the universal set of participants), and
- **I**∩**T** – All participants with both Instagram and Twitter accounts (i.e. the intersection between **I** and **T**).

3.2 Participant Demographics

We report the participant demographics across all four subsets in Table 1. We note that among all participants, 17.1% gave access to both their Twitter and Instagram accounts – leaving a majority (82.9%) of the participants unique to one of the two social media platforms. In terms of age, 80.6%-84.9% of the participants across all subsets were between 18-23. The Twitter subset has a slightly younger age distribution in comparison to the Instagram subset. As for sex, 75.0%-78.0% of the participants across all subsets are female. While most of the statistics on sex are relatively stable across subsets, we note a slightly higher percentage of females on Instagram and that there were fewer people who decline to disclose their sex if they granted access to both of their social media accounts. Lastly, only 0.8%-1.3% of participants across all subsets declared their nationality as Filipino and one or more nationalities.

3.3 Personality Trait Score

We report descriptive statistics of the personality trait scores of all participants in Table 2 and visualize the score distributions in Figure 1. All trait score distributions are unimodal and approximately symmetric with skewness values > -0.40 and < 0.04. We also report the Cronbach's alpha values for each trait in order to how consistent our participants were answering the Big Five Inventory (i.e. internal consistency). The alpha values indicate good reliability for Extraversion and Neuroticism, acceptable reliability for Conscientiousness and Agreeableness, and questionable reliability for Openness. As for correlation coefficients, most values showed negligible correlation. When coefficients weren't negligible, values showed low correlations, such as with Agreeableness and Conscientiousness, Neuroticism and Conscientiousness, Neuroticism and Extraversion, and Neuroticism and Agreeableness.

[1] *Pagkatao* is Filipino for *personality*, while *ko* refers to one's self or the word *my*. Hence, *PagkataoKo* is a play on the popular dataset, *myPersonality*.

Demographics	I∪T	I	T	I∩T
Count	3,128	1,380	2,283	535
Age				
Mean	21.2	21.4	21.0	21.2
SD	3.9	3.7	3.9	3.5
Age range				
18-20	53.9%	49.3%	55.9%	50.1%
21-23	29.3%	31.3%	29.0%	33.3%
24-26	9.3%	10.7%	8.5%	9.5%
≥27	7.5%	8.8%	6.6%	7.1%
Sex				
Male	21.0%	20.0%	22.0%	22.6%
Female	76.1%	78.0%	75.0%	76.3%
Intersex	0.5%	0.3%	0.6%	0.4%
Declined[1]	2.4%	1.7%	2.5%	0.8%
Nationality				
Filipino	99.2%	99.1%	99.1%	98.7%
Mixed[2]	0.8%	0.9%	0.9%	1.3%

[1] Those who declined to disclose their biological sex.

[2] Those who were Filipinos and had one or more other nationalities.

Table 1: The demographic statistics across all four subsets of participants: the universal set of all participants (**I**∪**T**), the set of participants with Instagram accounts (**I**), the set of participants with Twitter accounts (**T**), and the set of participants with both Instagram and Twitter accounts (**I**∩**T**).

3.4 General Data Characteristics

We summarize general statistics of user-generated data (e.g. posts, profile pictures) and account-related metadata (e.g. number of followers / following) across each of subgroup of participants in Table 3.

For the Instagram subset, the distribution of total collected posts per user is positively skewed with 72% of the subset having a total post count less than the mean (i.e. < 149.55) and 89% of the subset having fewer than one standard deviation plus the mean (i.e. < 377.03). There are 71 Instagram users with 0 posts and 75 Instagram users with total posts more than two standard deviations plus the mean (i.e. > 597.55). As for the posts themselves, only 83% of all collected posts have an image and 91% of the posts have a caption. Ideally, each post should have an associated image as one cannot post on Instagram without an image; however, we incurred a 17% loss in collectable image data due to the image link download issue discussed in Section 2.2. This issue also explains the missing 350 profile pictures. As for the missing 9% of posts without captions, we note that Instagram treats captions as an optional field when posting; hence, these posts really did not contain any text data. As for account-related data, we were only able to collect the total account recorded posts and the user's following count. Instagram's Legacy API was in the process of depreciating at the time of collection and did not allow for other metrics to be collected.

As for the Twitter subset, the distribution of total collected tweets per user is bimodal with roughly 49% of the subset falling between the 2200-3100 tweet count range and roughly 22% of the subset falling between the 0-600 range. There are 28 users with 0 tweets and 32 users with 3100-3200 tweets. Unlike posts on Instagram, all tweets contain text data. However, similar to the case with the Instagram subset, 804 users are missing a profile picture due to the image link download issue. As for account-related data, we were able to collect and report the total account recorded posts, following count, followers count, and favorite count.

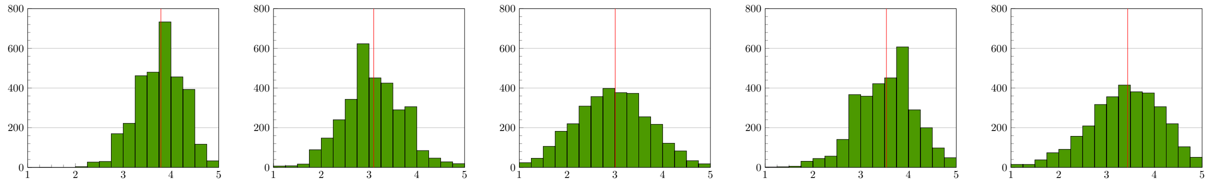
Of the total $3,128$ participants, only 17% of the participants granted access to both their Instagram and Twitter accounts. This subset retained 39% and 25% of the Instagram and Twitter posts, respectively. Despite the significant reduction in size, the subset's statistics are comparable to each platform's own subsets when factoring in that there are fewer outliers.

3.5 Temporal Characteristics of Post Data

We report the distribution of collected posts/tweets created per year for both social media platforms in Figure 2. As we collected as many posts/tweets as allowed, the Twitter data contains tweets made within an eleven-year period (2009 to 2020), while the Instagram data contains posts made within a ten-year period (2010 to 2020). The distribution of Instagram posts is approximately symmetric and peaks in 2016 ($n = 37,717$), while for Twitter, the distribution of tweets is left-skewed and peaks in 2019 ($n = 1,560,201$). Both distributions show a sharp drop off in 2020 due to the collection ending in Feb 2020.

Traits	Mean	SD	Alpha	Pearson Correlation Coefficient				
				O	**C**	**E**	**A**	**N**
O	3.7893	0.4837	0.6780	1.0000				
C	3.0984	0.6130	0.7844	0.1700	1.0000			
E	3.0066	0.7550	0.8249	0.1677	0.1379	1.0000		
A	3.5383	0.6075	0.7269	0.1309	0.2916	0.1907	1.0000	
N	3.4427	0.7462	0.8102	(0.1126)	(0.3695)	(0.2327)	(0.3016)	1.0000

Table 2: The mean, standard deviation, Cronbach's alpha, and Pearson correlation coefficients of each personality trait – **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism – with respect to all 3, 128 participants.



(a) Openness (b) Conscientiousness (c) Extraversion (d) Agreeableness (e) Neuroticism

Figure 1: Personality trait score distribution of all 3, 128 participants for each of the Big Five. The x-axis measures the raw trait scores, while the y-axis measures the number of individuals per bin. The red line represents the mean.

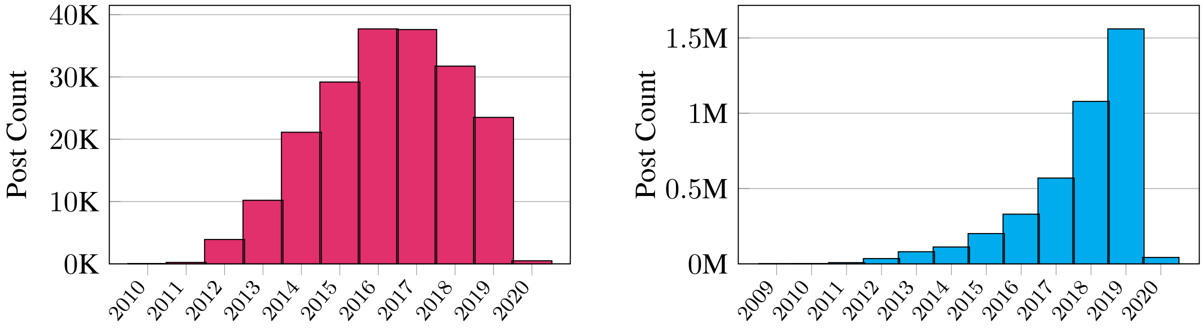


Figure 2: Histograms of dataset's posts made per year from Instagram (left) and Twitter (right)

3.6 Language Characteristics of Post Data

To highlight the multilingual aspect of our dataset, we observe language information at two levels: the entire post (*document-level*) and the individual words found in each post (*word-level*). In its raw form, the dataset only contains document-level language information for the Twitter data as this information was provided when collecting tweets via Twitter's API. To observe language information across the entire dataset, we process the text data to extract language information. We limit our observation to the following eight Philippine languages: *Bikol*, *Chavacano*, *Cebuano*, *English*, *Iloko*, *Kapampangan*, *Tagalog*, and *Waray*. While there are more Philippine languages than listed, these eight languages have some support from available language resources.

Data Pre-processing. For each individual post, we remove tokens with questionable or no language information, such as emojis, hashtags, usernames, URLs, and punctuation. We also lowercase all characters to reduce typical noise found in social media data. After the initial cleaning, we discarded all documents that were empty strings or contained only white space characters. This left us with 168, 723 posts from Instagram and 3, 979, 010 tweets from Twitter. We would like to note that pre-processing was done solely to extract

Platform	Data	Statistics	**I** $n = 1,380$	**T** $n = 2,283$	**I∩T** $n = 535$
Instagram	Collected Posts	Total	195,757	-	76,697
		Average	141.85	-	143.36
		SD	224.00	-	232.65
		Min / Max	0 / 1,902	-	0 / 1,902
		# w/ Caption	178,650	-	72,266
		# w/ Image	162,500	-	60,070
	Acct. Recorded Posts	Average	146.60	-	144.59
		SD	267.99	-	232.99
		Min / Max	0 / 5,680	-	0 / 1,902
	Following Count	Average	470.39	-	449.47
		SD	442.85	-	373.69
		Min / Max	0 / 6,338	-	0 / 4,427
	Profile Pictures	Total	1,030	-	385
Twitter	Collected Tweets	Total	-	4,018,628	1,033,089
		Average	-	1,760.24	1,931.01
		SD	-	1,016.71	987.62
		Min / Max	-	0 / 3,185	0 / 3,185
	Acct. Recorded Tweets	Average	-	8,003.57	9558.45
		SD	-	12,260.68	13,338.02
		Min / Max	-	0 / 162,738	0 / 103,381
	Following Count	Average	-	289.48	321.69
		SD	-	335.83	318.62
		Min / Max	-	0 / 7,079	0 / 3,501
	Followers Count	Average	-	333.93	320.58
		SD	-	1,023.46	393.51
		Min / Max	-	0 / 29,328	0 / 3,433
	Favorites Count	Average	-	8,517.47	8,886.03
		SD	-	12,252.05	10,734.24
		Min / Max	-	0 / 193,119	0 / 91,012
	Profile Pictures	Total	-	1,479	333

Table 3: Data characteristics of user-generated and account-related data across three participant subsets: all participants with Instagram accounts (**I**), all participants with Twitter accounts (**T**), and all participants with both Instagram and Twitter accounts (**I∩T**). The size of each subset (n) is also indicated.

language characteristics and that the raw data still contains this information.

Document-level Language Information. We extract document-level language tags using two language identifiers: *Polyglot* (Al-Rfou et al., 2013) and *FastText* (Joulin et al., 2016b; Joulin et al., 2016a). FastText supports all languages within our scope but forces a language tag even when it's uncertain. On the other hand, Polyglot covers all languages except Bikol, Chavacano, and Iloko and includes an *Undefined* tag when it lacks in confidence. Using the identifiers' output, we label a document based on the language tag with the highest confidence. When a language tag is outside of our scope, we assign the *Others* tag. Specific to Twitter data, we utilize the language metadata tag returned by Twitter's API, referred to as *Twitter Tag*, as a third language tag. The Twitter Tag only covers English and Tagalog and includes an undefined tag. After language extraction, we measure agreement among the assessed document-level tags through a *Majority Vote* (i.e. agreement $> 50\%$). If there is agreement among the tags, we assign the language tag. Otherwise, we assign a *Conflict* tag.

We summarize the results of our document-level language extraction in Table 4. The results show that English and Tagalog are the top two languages found on both platforms. English has a significantly higher usage on Instagram with a majority vote at almost 75.6% compared against the 47.9% majority vote on Twitter, while Tagalog has a significantly higher usage on Twitter (majority vote at 32.7%) versus that on Instagram (majority vote at 4.3%). As for the other Philippine languages, we note they occur significantly less often with Cebuano and Waray coming in third and fourth most used on both platforms. While this might indeed be true for the dataset, we take into consideration that different language identifiers do not align with each other – causing the majority vote to be low or result in zero. We also note that the extracted tags did not reach an agreement for 19.4% of the Instagram data and 17.0% of the Twitter data. Documents that fall under this category typically contain textspeak or some form of code switching between English and Tagalog.

We note that the Polyglot and FastText numbers are relatively similar despite the differences in their respective outputs; however, one glaring issue we would like to highlight is how Twitter Tag vastly differs from the two language identifiers. Twitter Tag indicates that there are 9.9% more Tagalog tweets than English, while numbers from Polyglot and FastText indicate that there are around two times more English than Tagalog tweets. Unfortunately, there are no specifics on how Twitter's language identifier works, but we speculate that Twitter uses different preprocessing techniques from our methods or uses information only accessible to Twitter.

To gain a better understanding of the issue, we performed a brief inspection of the Twitter documents. When agreement was reached, we note that 75% of the English tweets and 53% of the Tagalog tweets had perfect agreement across the three tags. These documents have a dominant language with respect to both the grammar and vocabulary. When there is disagreement between Twitter Tag and the other language identifiers, we note that it is rare ($< 0.05\%$ of the total tweets) for Twitter Tag to output English when Polyglot and FastText output Tagalog. On the other hand, almost 9% of the total tweets are labeled Tagalog by Twitter Tag when the the other two language identifiers agree on English. We observe that it is generally harder to determine these documents' language due to multiple factors, such as a balanced mix of words from both languages, multiple sentences following different grammar structures, and noise usually found in social media text. Based on our manual observation, we gained greater confidence in Twitter Tag particularly when it comes to the Tagalog labels. We also view Polyglot and FastText as sufficient off-the-shelf language identifiers but that they have a tendency to favor the English label. Hence, we caution interpreting the numbers too strictly and advise to keep in mind that the language characteristics of the data can be quite complex. Additionally, while these issues were solely observed on the Twitter data, we assert that the same issues with Polyglot and FastText may apply to the Instagram data but to a lesser extent.

Word-level Language Information. To observe word-level language information, we extracted the tokens and word types from our text data. We then compared how many tokens and vocabulary were found in a Philippine language word reference or dictionary. To serve as our reference, we used the words found in FastText's pre-trained word vectors (Grave et al., 2018) as there are resources for the languages within our scope except for Chavacano. We note that while FastText is a convenient resource, the word vectors' vocabularies are not unique from each other as they were trained on data from Wikipedia and CommonCrawl, which most likely included words from other languages.

We summarize the results of our word-level language information in Table 5. For the Twitter data, we note that the Tagalog word vectors provide the best coverage – providing vectors to 93.4% of our tokens, as well as 15.4% of the vocabulary. English comes in at a close second place covering 88.8% of the tokens and 15.1% of the

Language	**Instagram** ($n = 164,044$)			**Twitter** ($n = 3,870,153$)			
	PG	**FT**	**MV**	**PG**	**FT**	**TT**	**MV**
Bikol	-	0.00%	0.00%	-	0.01%	-	0.00%
Chavacano	-	0.01%	0.00%	-	0.01%	-	0.00%
Cebuano	0.48%	0.81%	0.12%	3.15%	2.99%	-	0.86%
English	82.56%	81.29%	75.58%	54.13%	54.15%	40.41%	47.94%
Iloko	-	0.04%	0.00%	-	0.22%	-	0.00%
Kapampangan	0.00%	0.00%	0.00%	0.00%	0.02%	-	0.00%
Tagalog	6.68%	5.88%	4.32%	26.89%	24.12%	50.26%	32.66%
Waray	0.40%	0.14%	0.01%	1.27%	0.77%	-	0.04%
Others	8.29%	11.82%	0.56%	13.96%	17.71%	6.26%	1.08%
Undefined	1.59%	-	0.00%	0.60%	-	3.06%	0.38%
Conflict	-	-	19.41%	-	-	-	17.01%

Table 4: The document-level language information of Instagram and Twitter documents. Language identifiers used were Ployglot (**PG**) and FastText (**FT**). Twitter's language metadata tag, referred to as Twitter Tag (**TT**), was also reported. Tag agreement was measured using a Majority Vote (**MV**) approach. Blank marks ('-') indicate the language identifier does not support the category.

vocabulary. As for Instagram, we are less certain of the word vector that provides the best coverage as English word vectors cover the vocabulary the most at 48.0% (versus Tagalog's 46.2% coverage), while the Tagalog word vectors cover the tokens the most at 94.7% (versus English's 94.6% coverage). As for the other Philippine languages, we note that the Waray word vectors come in a definitive third place both in terms of tokens and vocabulary coverage across the two platforms. After Waray, the ranking starts to vary across platforms. However, one observation we would like to point out is that the Cebuano word vectors have a higher token coverage (72.1%) on Twitter in comparison to the numbers of Iloko (68.1%) and Kapampangan (71.3%) despite having the second lowest vocabulary coverage at 2.9%. We speculate that the word vectors of Iloko and Kapampangan may have a sizeable overlap with other languages and that the higher token coverage may be an indicator that the Cebuano word vectors are able to capture a number of Cebuano function words.

4 Discussion and Future Directions

The PagkataoKo Dataset is a novel dataset for Filipino Automatic Personality Recognition that contains demographics, personality trait scores, and social media data from 3,128 Filipinos. The dataset is an improvement over the dataset of Tighe and Cheng (2018) having 12.5 times more participants and sourcing social media data from more than one platform. It has yet to be seen how much personality information is present in the dataset and how well personality models can compare against that of Tighe and Cheng (2018) and Tighe et al. (2020); however, solely based on the amount of data present, our current dataset provides a wider foundation to study how personality can manifest in the social media data of Filipinos – particularly as there are different forms of observable actions (e.g. text and image data from posts) and similar types of data expressed in different environments (e.g. language usage on Twitter and Instagram, profile picture usage on Twitter and Instagram).

While there is much potential in the dataset, there are a number of challenges that need to be carefully studied. First, descriptive statistics of the personality trait scores show that Openness has questionable reliability – raising the issue of whether or not the questionnaire is appropriately capturing the dimension. While collecting data using different personality instruments is indeed an option for future studies, we believe there may be additional insights that can be extracted by conducting APR by studying individual questionnaire item answers aside from the computed trait score. Second, the temporal characteristics of the posts show data spanning multiple years.

Platform	Token Count / Vocabulary Size		% Found in FastText Word Embeddings						
			BCL	CEB	EN	ILO	PAM	TL	WAR
Instagram	Tokens	1,786,309	74.8%	77.0%	94.6%	76.6%	80.5%	94.7%	82.0%
	Vocab	92,030	11.5%	11.9%	48.0%	13.3%	15.2%	46.2%	18.2%
Twitter	Tokens	33,451,199	68.1%	72.1%	88.8%	68.1%	71.3%	93.4%	75.0%
	Vocab	719,927	2.6%	2.9%	15.1%	3.2%	3.1%	15.4%	4.4%

Table 5: The percentage of tokens and vocabulary of each platform found in FastText's pre-trained word vectors across Bikol (**BCL**), Cebuano (**CEB**), English (**EN**), Iloko (**ILO**), Kapampangan (**PAM**), Tagalog (**TL**), and Waray (**WAR**). The total number of tokens and vocabulary size per platform are also indicated.

While personality traits are known to be relatively enduring across time (Larsen and Buss, 2008), traits aren't immune to change even through adulthood (Roberts and Mroczek, 2008). Additionally, Arnoux et al. (2017) was able to show that there is merit in exploring a shorter number of documents for personality prediction; however, their work mainly focused on English data. We speculate that more data might be needed to account for the noise brought about by code-switching found in our dataset. Hence, we encourage future work on APR to explore how the recency of one's posts might have an effect on APR prediction models. Third, the data characteristics show that there are a number of participants with either zero or a very low number of data points. Coupled with missing image data and, specific to Instagram, missing text data, future studies in Filipino APR would need to conduct experiments to find appropriate thresholds that determine when there's enough data to analyze one's personality and/or design a framework for APR that can handle missing data. Lastly, the language characteristics of the post data show that future studies working on the PagkataoKo dataset should primarily focus on extracting information from Tagalog and English text data because these languages were the most prevalent. While it would be of particular interest to study manifestations of personality in the other Philippine languages, the collection methods did not result in a sizeable amount of text data to study the other Philippine languages. Regardless, the nature of how Filipinos write on social media poses a serious challenge to text processing given multiple posts of a user could be in different languages or contain code-switching. We encourage future Filipino APR studies to focus on experimenting with methods to handle multilingual data – whether through combining language resources or by exploring language-independent approaches.

Acknowledgments

We would like to express our gratitude to the University Research Coordination Office of De La Salle University for supporting our research project (project number 52FU2TAY18-2TAY19). We would also like to thank the anonymous reviewers for spending their time to provide us with useful feedback.

References

Glenn Abastillas. 2018. You are what you tweet: A divergence in code-switching practices in cebuano and english speakers in philippines. In *Language and Literature in a Glocal World*, pages 77–97. Springer.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pages 1–16.

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Pilar Caparas and Leah Gustilo. 2017. Communicative aspects of multilingual code switching in computer-mediated communication. *Indonesian Journal of Applied Linguistics*, 7(2):349–359.

Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*, pages 7–10.

Rui Gao, Bibo Hao, Shuotian Bai, Lin Li, Ang Li, and Tingshao Zhu. 2013. Improving user profile with personality traits predicted from social media content. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 355–358.

Alastair Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24.

Alastair Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? personality, topic and motivation in blogs. In *Third International AAAI Conference on Weblogs and Social Media*.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011a. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156. IEEE.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011b. Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 253–262.

Lewis Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, 2(1):141–165.

Samuel Gosling, Adam Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Sharath Chandra Guntuku, Lin Qiu, Sujoy Roy, Weisi Lin, and Vinit Jakhetiya. 2015. Do others perceive you as you want them to? modeling personality based on selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 21–26.

Oliver John, Eileen Donahue, and Robert Kentle. 1991. The big five inventory–versions 4a and 54.

Oliver John, Laura Naumann, and Christopher Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research*, 3(2):114–158.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Simon Kemp. 2021. Digital 2021: Global overview report. *DataReportal*.

Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. 2014. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95(3):357–380.

Randy Larsen and David Buss. 2008. *Personality: Domains of Knowledge About Human Nature*. Boston: McGraw-Hill.

Alixe Lay and Bruce Ferwerda. 2018. Predicting users' personality based on their 'liked' images on instagram. In *The 23rd International on Intelligent User Interfaces, March 7-11, 2018*. CEUR-WS.

Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Tenth International AAAI Conference on Web and Social Media*.

François Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David Stillwell. 2013. Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

Scott Nowson and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167. Palo Alto, CA.

Scott Nowson and Jon Oberlander. 2007. Identifying more bloggers. *Proceedings of ICWSM*.

Veronica Ong, Anneke DS Rahmanto, Derwin Suhartono, Aryo E Nugroho, Esther W Andangsari, Muhamad N Suprayogi, et al. 2017. Personality prediction based on twitter information in bahasa indonesia. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 367–372. IEEE.

Gregory Park, H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934.

James Pennebaker and Laura King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 180–185. IEEE.

Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Brent Roberts and Daniel Mroczek. 2008. Personality trait change in adulthood. *Current Directions in Psychological Science*, 17(1):31–35.

Zahra Riahi Samani, Sharath Chandra Guntuku, Mohsen Ebrahimi Moghaddam, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2018. Cross-platform and cross-interaction study of user personality based on images on twitter and flickr. *PloS one*, 13(7).

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9).

Cristina Segalin, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri. 2017. What your facebook profile picture reveals about your personality. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 460–468.

Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing social media cues: Personality prediction from twitter and instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 107–108. International World Wide Web Conferences Steering Committee.

Yla Tausczik and James Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Edward Tighe and Charibeth Cheng. 2018. Modeling personality traits of filipino twitter users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 112–122.

Edward Tighe, Oya Aran, and Charibeth Cheng. 2020. Exploring neural network approaches in automatic personality recognition of filipino twitter users. In *Proceedings of the 20th Philippine Computing Science Congress*, pages 137–145.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Randall Wald, Taghi Khoshgoftaar, and Chris Sumner. 2012. Machine prediction of personality from facebook profiles. In *2012 IEEE 13th International Conference on Information Reuse & Integration*, pages 109–115. IEEE.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.

Novelty Detection in Community Question Answering Forums

Tirthankar Ghosal
ÚFAL, MFF
Charles University, Czech Republic
ghosal@ufal.mff.cuni.cz

Vignesh Edithal
IIT Patna
Patna, India
edithal.cs14@iitp.ac.in

Tanik Saikh
IIT Patna
Patna, India
1821CS08@iitp.ac.in

Saprativa Bhattacharjee
Government Polytechnic Daman
Daman, India
saprativa.bhatt@gov.in

Asif Ekbal
IIT Patna
Patna, India
asif@iitp.ac.in

Pushpak Bhattacharyya
IIT Bombay
Mumbai, India
pb@cse.iitb.ac.in

Abstract

Community Question Answering (CQA) forums are a popular place for open-ended question-answers and discussions by the public. Popular question answering sites have become a one-stop solution for the diverse information-seeking needs of the netizens. However, it is quite common that users sometimes ask the same question which may have been posted before and appropriately answered by the community members. It is not uncommon for users to perform an improper search and pose the same question again. Community members start responding to that question; however, the answer may already have been proposed previously on a different thread. When two questions are getting at the same problem, community members (moderators) in the forum often flag one as a duplicate of the other to help route traffic to high-quality *novel* questions and their correct answers. However, they usually do this manually based on the novelty or redundancy of the question. In this work, we try to mitigate this problem of detecting semantically equivalent *non-novel* questions automatically and flagging those. We also propose an approach to identify *novel* questions from CQA forums so that only the novel questions and corresponding answers threads stay, and semantic duplicates are removed. We make use of a Dynamic Memory Network (DMN) to assimilate information from multiple source questions to answer whether a new question is *novel* or a *semantically equivalent* question already exists. We introduce a new dataset for semantic-level novelty detection on community question answering. Our proposed approach attains performance improvements of +6.64% in terms of accuracy and +9.09% in terms of averaged F1 score over recent textual novelty detection methods. We would make our newly created dataset and the proposed approach available at `https://github.com/edithal-14/DMN-Novelty`.

1 Introduction

Community Question Answering forums are a convenient source of information for web users. Users post their questions on the forum, and fellow community members help them with pointers or answers to their diverse information needs. However, with the rapid growth of content within those websites, *redundancy* has become a problem. Unaware of existing solutions, users pose questions on a new thread that may already have been resolved in an existing thread. Such activity is widespread, and unaware community members start interacting on the new thread with solutions/pointers that may have already been discussed on an older thread. Usually, CQA forums have moderators who flag such *non-novel* questions and route the users to the existing solutions on the forum. To channel traffic to high quality and *novel* questions and answers on the forum, it is important to weed out *redundancies* and *non-novelties*.

Most of the CQA platforms request the users to go through the previously asked questions before posting a new question and post their question only if a similar question does not exist. But expecting such due diligence from each and every user is a tall ask in itself. Moreover, manually tagging questions as duplicates also requires a lot of effort on the moderators. This necessitates automatic techniques for efficient identification of non-novel or duplicate questions. Such a system could be used to tag and merge the duplicates out of the already asked questions and alert the users to the existence of a potential duplicate question while attempting to post a new question.

We leverage the DMN+ model (Xiong et al., 2016a) which proposed modifications to the *input* module and *memory* module of the original DMN framework (Kumar et al., 2016). Also, we extend the usage of DMN from word embeddings to sentence embeddings. We use the Infersent sentence encoder (Conneau et al., 2017) trained on the semantically rich SNLI corpus using GloVe word vectors. Experimental results show significant improvements over two deep learning-based baselines and two existing comparable systems. The major contributions of our current work are (i). An improved DMN framework for semantic level novelty detection in CQA forums, and (ii). A novelty detection dataset parsed from publicly available Stack Exchange (STE) data dump.

The remainder of the paper is organized as follows: Section 2 describes the related works. We describe the dataset in detail in Section 3. Methodologies adopted in this article are described in Section 4. System evaluation results obtained along with comparisons and rigorous error analysis are presented in Section 5. Finally, we conclude this article with some future research directions in Section 6.

2 Related Works

The problem of novelty mining is a long-standing problem in Information Retrieval (IR). The task has matured through several shared tasks, workshops, etc. Starting from Wayne (1997) to the novelty detection track as a part of The Text Retrieval Conference (TREC) workshop organized by NIST in the year of 2002 (Voorhees, 2002), 2003 (Voorhees, 2003) and 2004 (Clarke et al., 2004). Allan et al. (2003) investigated the tasks defined in the TREC 2020 novelty track, i.e., given a topic and list of documents relevant to the topic. The task is to first find the relevant sentences from the collection of documents and then find the novel sentences from the collection of relevant sentences.

The task of novelty detection can also be performed by Textual Entailment (TE). This idea had taken shape through the shared tasks organized in the year 2006 (Bentivogli, 2010), and 2007 (Bentivogli, 2011). In this era of deep learning, the availability of high-quality benchmark data has been the key bottleneck in advancing the novelty detection field. Ghosal et al. (2018b) first came up with a considerable amount of data, namely *Document Level Novelty Detection (DLND) TAP-DLND 1.0* and later extended version *TAP-DLND 2.0* to feed data-hungry deep neural models and adapted the novelty detection task from sentence level to document level.

Duplicate questions detection is a sub-task in QA. A pair of questions are assumed to be similar if both the questions can be satisfied with the same answer. It is a challenging task in two aspects: *viz. (i). There are many ways of asking a question, i.e., (question paraphrasing)* and *(ii). Asking a question has an implicit purpose, so even if two different questions seem like looking for the same solution, they can have entirely different purposes*. Being able to detect such questions leads to an increase in the accuracy of a QA system. Bogdanova et al. (2015) defined two questions can be considered as duplicates if the same answer can answer them. Robertson et al. (1994) considered two questions as a bag of words and computes scores between them. Further, a weighted matching (Inverse Document Frequency) between the tokens is performed for determining if two questions are close. The system proposed in Prabowo and Budi Herwanto (2019) detects duplicate questions in QA Website. The proposed system is equipped with GloVe pre-trained word embeddings, Convolutional Neural Network (CNN), and siamese network. Labeled data is precious, Rücklé et al. (2019) tried to mitigate this problem. They framed this problem into a zero-shot setting.

In contrast to the prior work, we frame the problem of identifying duplicate questions from a different viewpoint. We employ a novelty detection approach for this. There are almost no studies that address this problem from a novelty perspective. We consider duplication as an opposite characteristic of novelty. So given a pair of input questions, novelty detection system predicts as *novel* (non-duplicate) or *non-novel* (duplicate). With this intuition, we carry out the experiments performed in this article. Our model is based on the Dynamic Memory Network (DMN) (Kumar et al., 2016) technique. DMN has a special property of having a *memory component* and *an attention mechanism*. This kind of network-aided technique has been used for QA, but we are unaware of any such methods that use DMN for novel question detection in online CQA forums.

3 Dataset Description

We create a new 'novel-question' detection dataset from the Stack Exchange family of websites. All of the community-contributed questions and answers, along with comments, upvotes, downvotes, tags, and other metadata from all these sites, are published publicly by the Stack Exchange network on the *Internet Archive*[1] regularly. As of this writing, the latest available data dump is from 06 June 2022. In the data dump, each question is connected to a set of related questions through the *PostLinks* entity. One of the attributes of this entity is the *LinkTypeId*, possible values of which are 1 and 3, depending on whether the present question is or is not a duplicate of the related question, respectively. This information comes from the act of marking questions as potential *duplicates* of related questions by the CQA forum moderators. The Stack Exchange (STE) novelty dataset is thus automatically created by extracting pairs of related questions, and the ground truth of novelty is established based on the value of the *LinkTypeId* for each such pair (NOVEL, NON-NOVEL). The dataset thus created spans 50 different topics with an average of 4312 question pairs for each topic resulting in 215667 question pairs. In Figure 1 we present the statistics for the top ten topics in the STE dataset.

4 Methodology

Our proposed novelty-detection method is based on Dynamic Memory Networks (DMN), which proved to be very effective in question answering.

4.1 Dataset Used

SNLI: SNLI (Bowman et al., 2015) is a widely used, well-recognized Natural Language Inference (NLI) corpus. It has 570k human-written English example pairs that are labeled manually for balanced classification with the labels *entailment, contradiction, and neutral*. We use SNLI to train our sentence encodings method (Infersent).

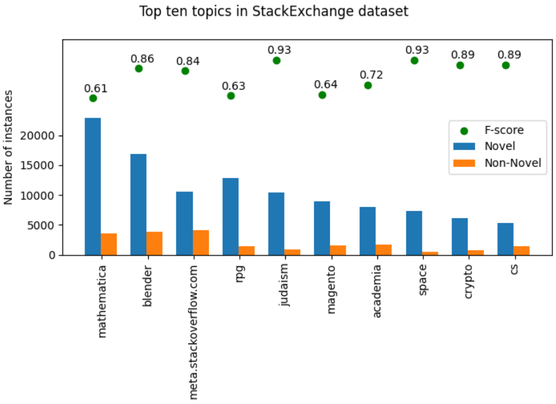


Figure 1: Statistics of top 10 topics by number of instances in the Stack Exchange dataset. The F-score corresponds to the performance of our DMN+ based model on that topic.

Stack Exchange Dataset: We parse the Stack Exchange data dump to create a Novelty detection dataset as described in Section 3 and run our proposed model on this newly created dataset.

4.2 Data pre-processing

We convert the source and target documents into a document matrix by encoding each sentence in the question into a 2048 dimensional vector. We employ *InferSent*(Conneau et al., 2017) for this purpose.

4.3 An improved DMN framework: DMN+

The basic DMN framework (Kumar and Irsoy, 2015) was proposed for dealing with QA problems. DMN consists of four independent modules *viz.* Input Module, Question Module, Episodic Memory Module, and Answer Module that can be improved independently. We re-implement the DMN+ model (Xiong et al., 2016b) for our task, which proposes to change the Input and Memory module of the original DMN framework.

Input Module: This module is responsible for encoding the inputs and adding contextual information to the inputs to produce fact vectors (f) used for further processing. The vanilla DMN model works on

[1] https://archive.org/details/stackexchange

the word level. In contrast, in the DMN+ model, we work on document level. We compute all sentences (contained in a document) embedding beforehand (using Infersent) and pass them through a Bi-GRU, and consider the output at each time-step is one fact. Let us consider our input consists of n sentences, and representation for each sentence from Infersent as $S_i, i = 1....n$; so the input is a sequence of n such representations as follows: $S = (s_1, s_2, s_3......s_n)$. This sequence is fed into a Bi-GRU, At each time step t, the network updates its hidden state $h_t = BiGRU(M[w_t], h_{t-1})$, where M is the embedding matrix and w_t is the sentence index of the t^{th} sentence of the input sequence. This module outputs the hidden states of the recurrent network. So we obtain f_n facts representation. The DMN+ approach solves two problems: (i). it allows for direct interaction between sentences that might be related to each other; at the word level, this type of interaction is difficult to capture, and (ii). a Bi-GRU allows for incorporating more contextual information from preceding and succeeding sentences, improving the quality of the fact vectors. We opt for Bi-GRU as it is lightweight, has fewer parameters, requires fewer computational resources, and is much faster than LSTM.

Question Module: We encode the target document's sentence representations via a Bi-GRU. Given the target question of T_Q sentences, hidden states for the question encoder at time t is given by $q_t = Bi - GRU(M[w_t^Q], q_{t1})$, M represents the embedding matrix as in the Input Module and w_t^Q represents the sentence index of the t^{th} sentence in the question. The word embedding matrix is shared across the input and question modules. This module produces the final hidden state of the Bi-GRU encoder, $q = q_{T_Q}$.

Memory Module: The outputs from the previous two modules are fed into the memory module. We attend the facts with respect to the question vector and the memory state (m) (initial memory state is the question vector itself) using a modified GRU called Attention GRU to create a context vector (c). The vanilla model sets the next memory state equal to this context vector, however the DMN+ model uses a memory update step to compute the next memory state using a ReLU layer over c, Q, and m. The final memory state after a defined number of memory state updates (also known as episodes or hops) is sent to the answer module. The final memory state should have enough information to answer the question after multiple attention passes over the incoming facts.

Attention GRU: The *attention GRU* is a traditional GRU with its update gate modified, which represents the importance of the incoming fact at the current time step. Following is a mathematical representation of a traditional GRU (here, x is the incoming fact, h is the hidden state of the GRU, i is the current time step, and • symbolizes the element-wise product)

$$u_i = \sigma(W^{(u)}x_i + U^{(u)}h_{i-1} + b^{(u)}) \quad (1)$$

$$r_i = \sigma(W^{(r)}x_i + U^{(r)}h_{i-1} + b^{(r)}) \quad (2)$$

$$\tilde{h_i} = tanh(Wx_i + r_i \bullet Uh_{i-1} + b^{(h)}) \quad (3)$$

$$h_i = u_i \bullet \tilde{h_i} + (1 - u_i) \bullet h_{i-1} \quad (4)$$

We replace the update gate u with an attention gate g as follows (here, f is the incoming fact, q is the question vector, m is the memory state, i is the current time step and t is the current memory update step)

$$z_i^t = [f_i \bullet q; f_i \bullet m^{t-1}; |f_i - q|; |f_i - m^{t-1}|] \quad (5)$$

$$Z_i^t = W^{(2)}tanh(W^{(1)}z_i^t + b^{(1)}) + b^{(2)} \quad (6)$$

$$g_i^t = \frac{exp(Z_i^t)}{\Sigma_{k=1}^{n} exp(Z_k^t)} \quad (7)$$

$$h_i = g_i^t \bullet \tilde{h_i} + (1 - g_i^t) \bullet h_{(i-1)} \quad (8)$$

Note that g is the output of a softmax layer which is essentially a probability distribution of how important each feature of the incoming fact is. The final hidden state of the Attention GRU is called the context vector (c).

Memory update step: This step is used only in the DMN+ model to compute the memory state based on the context vector. Note that the initial memory state m^0 is the question vector itself.

$$m^t = ReLU(W^t[m^{t-1}; c^t; q] + b) \quad (9)$$

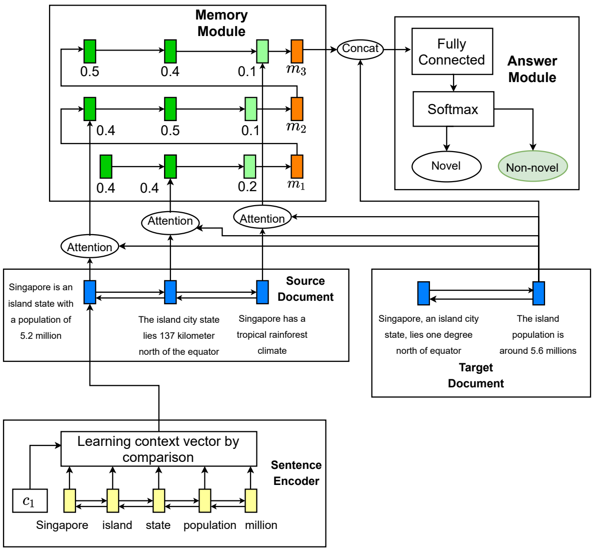

Memory Module
0.5
0.4
0.1
m_3
0.4
0.5
0.1
m_2
0.4
0.4
0.2
m_1
Concat
Fully Connected
Softmax
Answer Module
Novel
Non-novel
Attention
Attention
Attention
Singapore is an island state with a population of 5.2 million
The island city state lies 137 kilometer north of the equator
Singapore has a tropical rainforest climate
Source Document
Singapore, an island city state, lies one degree north of equator
The island population is around 5.6 millions
Target Document
Learning context vector by comparison
Sentence Encoder
c_1
Singapore
island
state
population
million

Figure 2: Novelty detection model based on DMN+ framework, colored blocks represent various tensors (e.g word/sentence embedding, attention and memory tensors).

Answer Module: This module differs according to the task at hand. We can use a linear layer with softmax for classification or one-word answer problems. We can use an RNN-based decoder network; if the answer is expected to be a sentence. We use the former approach, as our task is to classify the target document (question) as 'Novel' or 'Non-Novel.' The final memory state is concatenated with the question vector and passed through a ReLU layer followed by a softmax layer for classifying the target document as *novel* or *non-novel*.

5 Evaluation

5.1 Training

We use the Pytorch library for the implementation of the proposed model. We use the cross-entropy loss function since we have the softmax layer as the final layer to classify into Novel or Non-Novel classes. We initialize the model parameters using the Xavier initialization method, and we optimize them using Adam optimizer (Kingma and Ba, 2017). We decrease the learning rate (LR) by a factor of 10 on the validation accuracy plateau with the patience of 3 epochs (if validation accuracy does not improve by 1% in 3 epochs); this allows the optimizer to escape the local minima when it is stuck in one and prevents the model from over-fitting. We use a batch size of 32. We train the model for 25 epochs with an early stopping of 10 epochs (stop if the validation accuracy does not improve in 10 epochs). We chose the model with the best validation accuracy for testing. We perform model hyper-parameters tuning manually. The original DMN+ model uses three memory update iterations (hops) in its memory module; however, we observe that four hops provides the optimal result, further increasing the number of hops reduces the accuracy.

5.2 Results

We test our model on the newly introduced Stack Exchange (STE) novelty dataset. We compare the results with two deep learning-based models that we consider as the baselines and two other existing comparable systems, *viz.* (i) RDV-CNN (Ghosal et al., 2018a) and (ii) Decomposable attention-based model (Ghosal and Edithal, 2020). For each of the 50 topics in the dataset, we split the available document pairs into 80:20 ratios for training and testing, respectively. We then individually train and evaluate the model on each topic separately. Finally, we average the performance of the model across all the topics. We show a comparison of the results in *Table 1*. From the Table we conclude that leveraging the DMN+ framework to obtain a joint representation of a pair of source and target documents and then using it for Novelty judgment provides much better accuracy than simply passing the sentence embeddings through a BiLSTM.

5.3 Analysis

We now show an example **non-novel** (redundant) source/target document pair, where our proposed model can capture document redundancy. In contrast, its close competitor, the decomposable attention model (Ghosal and Edithal, 2020) fails to detect the redundancy and classifies the document pair incorrectly as Novel. We present the model's predictions along with a heatmap to explain the predictions in *Figure 3*.

Source **[s1]** Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. **[s2]** Most people who fall sick with COVID-19 will experience mild to moderate symp-

Model	Accuracy	P (N)	R (N)	P (NN)	R (NN)	F1 (avg)
Doc2Vec + BiLSTM + MLP	64.88%	0.66	0.62	0.64	0.67	0.65
Inner attention + BiLSTM + MLP	70%	0.60	0.70	0.75	0.75	0.70
RDV-CNN (Ghosal et al., 2018a)	72%	0.71	0.61	0.70	0.71	0.72
Decomposable Attention (Ghosal and Edithal, 2020)	77%	0.82	0.68	0.73	0.85	0.77
Inner attention + DMN+ + MLP	**83.64%**	0.86	0.80	0.81	0.87	**0.84**

Table 1: Comparing our model with two baselines and two comparing systems (RDV-CNN and Decomposable attention) on the STE dataset. Average performance across all the 50 topics in the STE dataset. N→Novel, NN→Non-Novel

Topic	Link to Source Document	Target Document	Gold class	Predicted class	Comments
blender /08761.json	https://blender .stackexchange.com /questions/16267/i-dont-know-how-i-locked-view-offset-but-how-do-i-unlock-it	https://blender .stackexchange.com /questions/47935/how-to-remove-revert-blender-object-centre-view	Novel	Non-Novel	Domain specific parlance and overlapping named entities caused the model to predict Non-Novel
cs /00632.json	https://cs.stackexchange.com/questions/4800/the-order-of-growth-analysis-for-simple-loop	https://cs.stackexchange.com/questions/10813/decreasing-runs-of-inner-loop-in-outer-loop	Non-Novel	Novel	Inability of the model to understand mathematical formatting and programming language syntax causes wrong prediction
academia /00074.json	https://academia .stackexchange.com /questions/1190/what-are-some-good-project-management-tools-for-academics	https://academia .stackexchange.com /questions/1273/use-cases-of-org-mode-as-a-scientific-productivity-tool-for-academics-without-pr	Non-Novel	Novel	There are instance of incorrect gold labels in this dataset, since this dataset is created by an algorithm which uses linked posts and post metadata, it is prone to mistakes. In this case the given document pair is clearly Novel, however, the gold label states Non-Novel

Table 2: Error analysis of the DMN+ model using STE instances, topic corresponds to the topic specific Stack Exchange forum where the questions were asked. The comment column explains the cause of the miss-classification.

toms and recover without special treatment. **[s3]** The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. **[s4]** These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces. **[s5]** You can be infected by breathing in the virus if you are within close proximity of someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose or mouth. **[s6]** You can reduce your chances of being infected or spreading COVID-19 by regularly and thoroughly cleaning your hands with an alcohol based hand rub or wash them with soap and water. **[s7]** Washing your hands with soap and water or using alcohol based hand rub kills viruses that may be on your hands.

Target (Non-Novel) **[t1]** COVID-19 symptoms are usually mild and begin gradually. **[t2]** Some people become infected but don't develop any symptoms and don't feel unwell. **[t3]** Most people (about 80%) recover from the disease without needing special treatment. **[t4]** Older people, and those with underlying medical problems like high blood pressure, heart problems or diabetes, are more likely to develop serious illness.

5.4 Error Analysis

In *Table 2* we present a few instances from the STE dataset wherein our model failed to classify the novelty of the document pair correctly. In some cases, our model predicts a pair of source and target documents as non-novel or duplicate due to a signif-

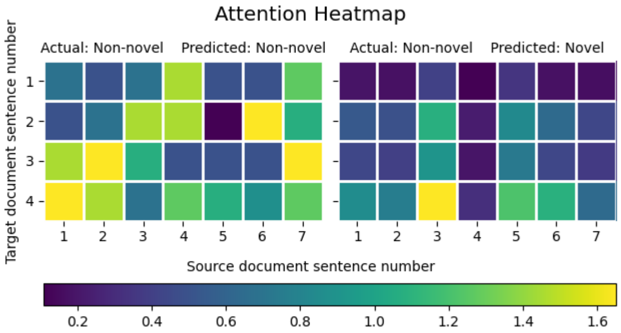


Figure 3: Heatmap denoting attention values between source and target sentences for the document pair mentioned in the Analysis section. *Left:* attention values from the last hop of the episodic memory module in our DMN based model. *Right:* attention values from the sentence comparison step of the Decomposable attention model.

icant overlap of named entities between the two. For instance, consider the pair for the topic *blender*, our model incorrectly predicts it as a non-novel pair mainly because of the use of domain-specific terminologies and due to the occurrence of named entities such as *blender*, *numpad*, *view* and others in both the source and the target.

In some other cases, the model makes wrong predictions due to the inability to understand the mathematical formatting and programming language syntax used in the questions' text. An example of such a pair is the one under the topic of *cs*, in which both the source and target consist of code snippets and inline mathematical expressions. Moreover, there are also instances of erroneous duplicate tagging of the question pairs in the gold set. This might be attributable to the subjective nature of the job, as often in CQA forums, the questions marked as duplicates and merged by moderators are later unmerged after reviewing the appeals from the original user who posted the question or even fellow community members. This is illustrated by the pair from the topic *academia*.

6 Conclusion and Future Work

In this work, we address the problem of duplicate question identification in community question-answering forums from the perspective of textual novelty detection. To the best of our knowledge, no prior work has leveraged textual novelty detection for tackling this problem. We use a deep Dynamic Memory Network, specifically the DMN+ for assimilating information from multiple source questions to detect the novelty of a target question at the semantic-level. Our method outperforms the deep learning-based baselines and recently proposed textual novelty detection methods. We also propose a new dataset consisting of 215K novel and non-novel question pairs over 50 different topics. We automatically create this dataset from the publicly available data dumps of the Stack Exchange network of websites.

In the future, we would like to validate our approach on other CQA forums such as Reddit and Quora questions/posts. We would also like to investigate how the recent large contextual language models would perform for this problem. We envisage that our novel investigation of associating textual novelty detection to detect semantic duplicates on the web would aid in several downstream tasks to alleviate the quality of information available.

References

James Allan, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, page 314–321, New York, NY, USA. Association for Computing Machinery.

Magnini B. Dagan I. Dang H.T. Giampiccolo D. Bentivogli, L. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC 2010), November 15-16, 2010 National Institute of Standards and Technology Gaithersburg, Maryland, USA.*

Clark P. Dagan I. Dang H. T. Giampiccolo D. Bentivogli, L. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *In TAC 2011 Notebook Proceedings, November 14-15, 2011, Gaithersburg, Maryland, USA.*

Dasha Bogdanova, Cícero dos Santos, Luciano Barbosa, and Bianca Zadrozny. 2015. Detecting semantically equivalent questions in online user forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 123–131, Beijing, China, July. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642.

Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the trec 2004 terabyte track. In *TREC*, volume 4, page 74.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Tirthankar Ghosal and Vignesh Edithal. 2020. Is your document novel? let attention guide you. an attention-based model for document-level novelty detection. apr.

Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018a. Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2802–2813, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. TAP-DLND 1.0 : A corpus for document level novelty detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Ankit Kumar and Ozan Irsoy. 2015. Ask me anything: Dynamic memory networks for natural language processing. arXiv:1506.07285. version 5.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1378–1387.

Damar Adi Prabowo and Guntur Budi Herwanto. 2019. Duplicate question detection in question answer website using convolutional neural network. In *2019 5th International Conference on Science and Technology (ICST)*, volume 1, pages 1–6.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019. Neural duplicate question detection without labeled training data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1607–1617, Hong Kong, China, November. Association for Computational Linguistics.

Ellen M Voorhees. 2002. Overview of trec 2002. In *Trec*.

Ellen M Voorhees. 2003. Overview of trec 2003. In *TREC*, pages 1–13.

Charles L Wayne. 1997. Topic detection and tracking (tdt). In *Workshop held at the University of Maryland on*, volume 27, page 28. Citeseer.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016a. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2397–2406. JMLR.org.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016b. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417.

A Graph-Based Approach Leveraging Posts and Reactions for Detecting Rumors on Online Social Media

Asimul Haque
Department of Computer Science
South Asian University
New Delhi-21, India
asimulhaque@gmail.com

Muhammad Abulaish, SMIEEE
Department of Computer Science
South Asian University
New Delhi-21, India
abulaish@sau.ac.in

Abstract

In this paper, we present a novel graph-based contextual and semantic learning approach for detecting rumors on online social media. The underlying hypothesis is that social media entities are intertwined, and if an event unfolds then similar narratives or user reactions with common interests get circulated. The proposed approach uses tweets and their reactions to understand the underlying interaction patterns and exploits the textual and latent information. Textual data is modeled as a words co-occurrence graph, which produces two prevalent categories of words – *substantial words* and *bridge words*. These words serve as building blocks for constructing contextual patterns for rumor detection by computing node-level statistical measures. The contextual patterns are further enriched by identifying negative emotions and inquisitive aspects in the reactions. The patterns are finally ranked and only top-k check-worthy patterns are used for feature generation. In order to preserve the semantic relations, we use word-level GloVe embedding trained over a Twitter dataset. The proposed approach is evaluated over a publicly available PHEME dataset, and compared with various baselines and SOTA techniques. The experimental results are promising and the proposed approach seems useful for rumor detection on online social media.

1 Introduction

The increasing popularity of online social media has motivated various actors to use them for spreading misinformation to set their personal agendas. Misinformation is dangerous, and it can stymie our efforts to address global challenges as many issues are being fuelled and distorted by it. Misinformation acts like a virus in the sense that it exploits our weaknesses, biases, prejudice, and emotions. It has deadly consequences, including polarizing debates, creation or deepening of societal tensions, undermining truth, and manipulation of the electoral processes. What we share online can have consequences in the real-world. The *world economic forum* has ranked massive digital misinformation as one of the top global risks[1]. The United Nations urged social media users to *"pause- take care before you share"* on the *world social media day* to combat misinformation. The world health organization termed misinformation as an *"infodemic"* which is spreading faster than COVID-19, disrupting public health efforts and distorting the sound scientific guidance[2].

Manual fact-checking sites like Snopes, PolitiFact, and FactCheck are available with some traditional media fact-checkers, such as (i) Reality Check – a fact-checking arm of the BBC which is reported in an article entitled *Coronavirus: the human cost of fake news in India*[3], and found spam propagating life-threatening consequences based on the false claims related to coronavirus outbreak, Delhi riots, citizenship amendment act, and claims about the minority com-

[1] https://www.weforum.org/reports/the-global-risks-report-2020

[2] https://www.who.int/health-topics/infodemic

[3] https://www.bbc.com/news/world-asia-india-53165436

munity, (ii) Verified, which is an United Nation's initiative providing facts and life-saving information to the citizens of the world, (iii) Google News Initiative, which provides funding to fight misinformation during the COVID-19 pandemic, (iv) NewsGuard, which is an internet tool for tracking misinformation. It investigated and flagged several Facebook pages as rumor spreaders. Apart from these, social media firms are also concerned with the identification of misinformation on their sites and consequently employing experts and encouraging users to flag or report posts that are not credible.

In the recent era of social media, a type of misinformation that spreads across the network in a short time span is known as a *rumor*. Rumor consists of controversial news, including factually incorrect information about celebrities, politics, crises, and social affairs. Information keeps circulating across social media platforms with unchecked veracity. In due course, the veracity of any information may be determined as true or false, or the same may remain unchecked (Zubiaga et al., 2018). Any kind of rumor or false news disseminates faster than authentic and true news. However, in the case of rumors on political matters, the rapidity of spread and the negative consequences of rumors or fake news is exacerbated (Vosoughi et al., 2018).

The ruckus caused by the spread of misinformation with nefarious motives distorts the truth and tarnishes the spirit of social media platforms. Therefore it is essential to curtail the spread of false rumors and elevate trust in the whole ecosystem of social media. Accordingly, it is becoming a challenge for social media scientists and researchers to devise effective algorithms for determining the veracity of any information on the breeding ground of social media. However, the methods devised for identifying and determining the veracity of rumors on social media, especially on Twitter, are based on hand-crafted features that come from two main aspects – content and users' social context (Zubiaga et al., 2017). Researchers are generally using traditional machine learning approaches for learned-representation of deep learning techniques to classify rumorous messages. There are recent works on determining the veracity of rumors, but very few of them focus on the detection of rumors based on the contextual representation learning using a graph-based approach. When an event begins to unfold, a similar type of user responses and a comparable set of patterns are generated. The re-circulation of the original post and showing skeptical or negative responses create echo chambers, convincing us to incorporate reactions and retweets while addressing the truthfulness of a tweet.

In this paper, we present a graph-based contextual learning representation for detecting rumors in Twitter. The graph-based approach is exploited to capture the contextual information from the tweets. Word co-occurrence graphs are constructed using textual data, and node metrics like *eigenvalue centrality* and *clustering coefficient* are computed. The *clustering coefficient* sets out the topical words that are generally clustered together, while *eigenvalue centrality* lists the words that connect the topical words. These two families of words are collaborated to create patterns. The novelty of our approach is in incorporating reactions with the source tweets to capture their inherent semantic affinity. Our approach is also unique in its way of graph-based representation learning and identifying two prevalent categories of words to extract the rumorous patterns. Moreover, the emotional and inquisitive words are considered to make patterns more generic. The ranking of patterns is performed through *tf-idf* weight score and top-k check-worthy patterns are extracted. For preserving the semantic relations, word-level GloVe embeddings learned from a Twitter dataset is applied. The tweets are split into n-grams, and *Cosine* similarity is used to generate the feature vectors by calculating the similarity between the patterns and tweets. Different classification models are trained over the original and a balanced distribution of the publicly available PHEME dataset. In comparison to the word-based method, the pattern-based approach has the advantage of being more representative and retaining the syntactic sense.

2 Related Works

In recent years, rumor and fake news detection are becoming one of the most explored areas of research. Although the rumor is an old phrase, the implications of such mis/disinformation on online social networks become apparent when events unfold as breaking news, prompting individuals to rely on social media for news and information. It can be said that the term fake news is also old, but got popularized after the 2016 presidential election in the USA. The very first work tackling the detection of emerging rumors was proposed in (Zhao et al., 2015), based on extracting correction and verification signals using regular expressions. Alternately, the authors in Zubiaga et al. (2017) leveraged the context using CRF with assumptions that context plays a crucial role in determining the rumors.

Gradually, research on the subject of rumor identification progressed, with mechanisms ranging from evaluating the rumorous nature of tweets to detecting stances and establishing the truth value of social media posts (Mohammad et al., 2016; Kochkina et al., 2017; Rosenfeld et al., 2020). The transformation of approaches varies from traditional models like SVM and Random Forest to probabilistic graphical models like Bayesian classifier and deep neural networks like LSTM and GCN (Castillo et al., 2011; Bai et al., 2021). Some researchers have explored behavioral, emotional, and sentimental aspects for the detection of rumors and fake news (Ajao et al., 2019). Abulaish et al. (2019) proposed a graph-based approach for rumor detection using `POS` tags to identify anxious and doubtful terms present in microblogging posts.

In recent years, various deep learning approaches are proposed for rumor detection. Some recent approaches are based on Graph Convolutional Neural Net (GCN) Bai et al. (2021) using a relationship between source and reply, whereas Dong et al. (2019) identified multiple rumor sources without knowing the underlying propagation structure. In Tu et al. (2021), the authors proposed a CNN-based model using text and propagation structure. Ma et al. (2019) applied a Generative Adversarial Network (GAN) based model to detect rumor-inductive patterns even for a low-frequent rumor. Nguyen et al. (2020) discussed a graph learning framework for fake news detection using the social structural engagements of the users. Compared with traditional textual content and social user features, the images and video-based rumors have been researched less. Deep learning has encouraged researchers to explore multimodal features of rumors (Zhou et al., 2020). Related research has looked into many elements of detecting rumors and fake news on social media, but they are less focused on inferring underlying information and extracting patterns from tweets and reactions. Moreover, of all the strategies produced, only a few are centered on graph-based representation learning mechanisms.

3 Proposed Approach

Following the existing state-of-the-art works, we consider rumor detection as a binary classification problem. Given a tweet $t_i \in T$ of a `Twitter` dataset T, rumor detection problem aims to estimate a function r which predicts the class level of t_i as a rumor or non-rumor. Mathematically, it can be defined as r: t_i $\rightarrow \{0, 1\}$ such that,

$$r(t_i) = \begin{cases} 1 & \text{if } t_i \text{ is a rumor} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

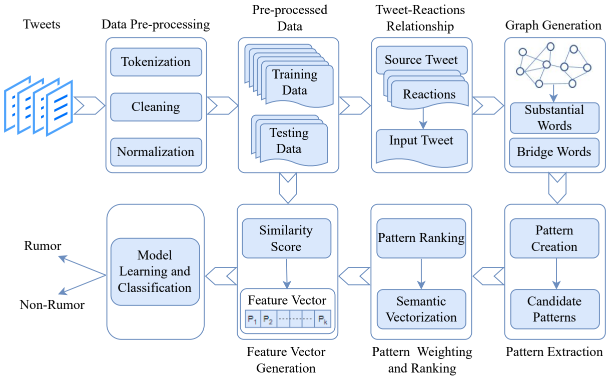


Figure 1: Work-flow of the proposed approach for rumor detection

The work-flow of the proposed approach for rumor detection is presented in figure 1. Different functionalities of the proposed approach are explained in the following sub-sections.

3.1 Data Pre-processing

In this step, several data pre-processing tasks such as tokenization, cleaning and normalization are performed on the `Twitter` dataset. All the tweets and reactions are tokenized with white space. To normalize the dataset and avoid any bias towards any events or twitter-name, we replace the twitter-specific tokens *hashtags, urls, retweet* and *mentions* with the tag ≺*hashtags*≻, ≺*urls*≻, ≺*retweet*≻ and ≺*mentions*≻, respectively. The punctuation and emoticons such as ?, !, and :(are not removed because they play a significant role in examining the writing style of rumors. Stop-words and capital letters are significantly important to understand the context of tweets. The capital letters show emphasizing and stop-words intact syntactic sense. Therefore, stop-words are also retained to aid in forming rumor patterns.

3.2 Inferencing Tweet-Reactions Relationships

This section aims to identify closeness between the source tweets and reactions. The reactions describe the underlying latent information. Users show *skepticism*, *correction*, and *verification* in their reactions towards the truthfulness of a tweet. Some users give opinions or clues regarding the factuality of a tweet. Therefore, tweets and reactions are highly correlated. The reactions provide supplements to the source tweet that enhance and describe their quality and are also helpful for identifying the underlying patterns in rumorous tweets.

For making tweets explainable towards rumors, the reactions are incorporated with them to get additional knowledge. When combining of tweets and reactions, the repetition of the same tweets in reactions should be avoided to make them unique and descriptive. For this purpose, the symmetric difference approach is used at the sentence level on the tweets and their reactions. For any tweet t_i and its set of reactions $R_{t_i} = \{r_1^{t_i}, r_2^{t_i}, ..., r_n^{t_i}\}$, the repetitions are removed using equation 2 by calculating the symmetric difference of a tweet and all its reactions.

$$SD_i = t_i \Delta \{r_n^{t_i}\}_1^n \tag{2}$$

The symmetric differences between a tweet and reactions can be formally defined as:

$$t_i \oplus \{r_n^{t_i}\}_1^n \tag{3}$$

To maintain the inclusiveness of a tweet and its reactions, a single input tweet is created by combining a tweet and its deduplicated reactions by making a union of all the symmetric differences SD_i. We define this relation in equation 4, where $X = (t_i, R_{t_i})$ is the set of a tweet and the union of symmetric differences SD_i of a tweet $t_i \in T$. This equation 4 represents an input tweet X_i in which the first sentence is a tweet followed by all its deduplicated reactions.

$$X_i = t_i + \bigcup_1^n SD_i^{t_i} \tag{4}$$

3.3 Graph Generation

With minimum domain knowledge, the graph-based method can effectively capture linguistic variation and contextual information in textual data. The input tweets are represented as a word co-occurrence graph, $G(V, E)$, where V is a finite set of nodes representing words and E is the set of edges representing the relationship between the nodes. To preserve the underlying structure of the input tweets, edges are defined as a sequence of words in an input tweet with a window size of 2. The graph is used to model two prevalent categories of the words – substantial words and bridge words that are identified by computing node-level statistical measures. The key objective is to collect the words that are relevant to constructing rumorous patterns. Two operations are performed on the graph for computing node statistics; *clustering coefficient* and *eigenvector centrality*. *Clustering coefficient* is used to identify substantial and topical words, whereas *eigenvector centrality* is used to identify bridge words that provide connections to the substantial words.

3.3.1 Substantial Words

When any event starts unfolding, a similar type of tweets start posted on social media that contains a similar set of responsive or reactive words. These words can be psychological, sentimental, or skeptical that are normally clustered together to impose sentiments, feelings, or suspicions about the event. These types of words are also connected with each other through the same words that we recognize as bridge words. Therefore, in order to trace the clustering behavior of such topical words, the *clustering coefficient* of each node $v_i \in V$ is calculated using equation 5.

$$C_{c(v_i)} = \frac{2|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)} \tag{5}$$

The average clustering coefficient of a node v_i is defined as $C_{c(v_i)}/|V|$ and the average clustering score of the graph G is defined as $\sum_{v \in V} C_{c_v}/|V|$. At the end, the list of substantial words is compiled with the words having a clustering coefficient value greater than the threshold θ_{C_c}, which is determined empirically as a value just greater than the average clustering score of the underlying graph.

3.3.2 Bridge Words

As the name suggests, this category of words makes a bridge between two substantial words and contributes to making rumorous patterns. It is used to deliberate the meaning of the substantial words and categorize the informal representation of the words because people use to write in casual ways on social media. To consider the frequent as well as focal nodes in the graph and based on the fact that a node is important if its neighbors are important, we choose the *eigenvector centrality* measure to identify bridge words. The benefit of choosing *eigenvector centrality* is to avoid frequent but irrelevant words in the graph. The *eigenvector centrality* of a node $v_i \in V$ is calculated using equation 6, where N_i is the neighbors of v_i and $A(i, j)$ is the adjacency matrix of the graph G. In other words, the centrality of a node v_i is proportional to the combined centrality values of its neighbors $v_j \in N_i$.

$$C_{e(v_i)} \propto \sum_{v_j \in N_i} A(i,j) C_e(v_j) \tag{6}$$

Now, it can be re-written in a matrix form as given in equation 8, where λ is a proportionality constant.

$$x \propto Ax \tag{7}$$

$$\lambda x = Ax \tag{8}$$

This equation looks exactly like the eigenvector equation. The centrality vector x is the eigenvector of the adjacency matrix $A(i, j)$ associated with the eigenvalue λ. By virtue of the *Perron-Frobenius* theorem, it takes the largest value of λ, and we find the corresponding eigenvector that is positive and unique, giving the *eigenvector centrality* of each node $v_i \in V$. The list of bridge words is compiled with the words having a centrality value greater than the empirically calculated threshold θ_{C_e} value.

3.4 Pattern Extraction

The objective of this step is the extraction of the rich rumorous patterns that are found more frequently in rumors. The order of words in any language is critical for delivering meaningful information. People communicate based on syntactic rules to convey the proper meaning. The same syntactic rule is followed in social media while writing posts. During the extraction of check-worthy patterns, we have considered the syntactic rule as well as the emotion and skeptical nature of social media users. We make the patterns as a combination of above-mentioned two categories of words. Keeping syntactic rules and extracting meaningful patterns; first, the nodes that satisfy the threshold condition are identified from both categories and located in the graph. Thereafter, we find the connections between the words of different categories. The connections are basically the edges between a node v_i and all its neighbor N_i. Since we are interested in considering the patterns that consists of maximum words to get more contextual and syntactic meanings, we have taken 3-word patterns that are created by combining a node with its two neighbors in a way that not all three nodes are from the same category. Finally, all possible n patterns are collected and stored in the list P such that $P = \{p_1, p_2, \ldots, p_n\}$.

3.4.1 Candidate Patterns Selection

The check-worthy patterns identified from list P are named as candidate patterns. The following steps explain the selection of the candidate patterns.

The first step is based on the assumption that the rumor contents are related to a specific event, i.e., when an event unfolds, individuals begin to circulate similar sets of responses and generate comparable types of information without verifying their veracity. Therefore, to incorporate event-specific patterns as well as making patterns close to being rumorous, the same set of procedures discussed above are applied to that portion of training data which is labeled as rumor only. Then this newly obtained list of patterns P_{new} is used to shorter the patterns list P. It is achieved through matching the list P with a newly designed list of patterns P_{new}. The matching is performed as described below:

N-gram matching: The *n*-gram matching works on the word level, providing the collection of patterns that match with rumor patterns. The matching is done at the trigrams level, i.e., the patterns that match all three words are collected in a list, P_R, named as rumor patterns list.

Similarity matching: In the above step, the exactly matched patterns are captured, but some similar patterns are left out that are extracted using a similarity measure. We have used Cosine similarity for this purpose. The Cosine similarity between a pair of patterns is calculated using equation 9, where p_i is the i^{th} pattern of list P and p_{new_j} is the j^{th} pattern of list P_{new}. The obtained patterns from this step are appended to the rumor patterns list P_R.

$$Cosine(p_i, p_{new_j}) = \frac{p_i \cdot p_{new_j}}{||p_i|| \cdot ||p_{new_j}||} \quad (9)$$

The first step gives the rumor patterns that are event-specific and about the content of the in-hand dataset. To make patterns generic, we consider the responsive behavior of social media users since the writing style and user behavior are mostly the same for any sort of unverified or rumorous social media posts. Incorporating reactions with tweets are advantageous to examining the emotional, correcting, and verifying contents in the input tweets.

The second step incorporates the sentimental and emotional words in the patterns list P. As discussed in Vosoughi et al. (2018), the reactions to rumors contain fear, surprise, and disgust. People re-circulate the rumorous tweet with negative sentiments. To consider the negative emotional words, `NRC Word-Emotion Association Lexicon` is used, consisting of words along with associative emotions *(anger, fear, anticipation, trust, surprise, sadness, joy*, and *disgust)* and *(negative* and *positive)* sentiments (Mohammad and Turney, 2013). The patterns having negative emotional words are extracted from the pattern list P and produced as a part of a list of emotional patterns P_E.

The third step considers the skeptical nature of social media users. When a tweet gets posted, people show skepticism and start *questioning* and *verifying* them in the form of support and denial. The correction and verification inquiry in the replies are also observed in (Zhao et al., 2015). Inspired by their work, the regular expressions such as $(true|not|true)$, $(real?|really?)$, $(rumor|debunk)$, and $(false|fake)$ are constructed and passed through the pattern list P. The patterns containing the skeptical words are extracted and considered as a part of skeptical patterns list P_S.

The above three steps extract the generic and specific patterns from the list P. Thus, candidate patterns have consolidated three categories of pattern lists – rumor patterns, emotional patterns, and skeptical patterns.

3.5 Pattern Weighting and Ranking

In this step, the top-k check-worthy patterns are selected from each list of candidate patterns. Since we are interested in finding highly rumorous patterns, a weight value is assigned to each pattern to define its ranking. It is achieved through *tf-idf* weight scoring, where each pattern of all three categories of the candidate list is assigned a weight score through the rumor training corpus tc. Furthermore, semantic vectorization is performed using embeddings to preserve semantic richness.

We modified the standard formula of *tf-idf* for calculating the weight score to map the patterns close to being rumorous. For this purpose, we split the input tweets of the rumor training corpus into n-grams where n is equal to the number of words in a pattern p_i. The modified *tf-idf* is applied separately on rumor patterns list P_R, emotional patterns list P_E, and skeptical patterns list P_S to assign weight scores to every pattern of the respective categories. For any of the above lists, each pattern of that list is considered as a term, and its frequency value is calculated, named as pattern frequency PF. We pass a pattern through the n-grams rumor corpus, and word co-occurrence is counted. The pattern frequency of a pattern PF_{p_i} is defined in equation 10, where, tc is the rumor training corpus, $f(p_i, tc)$ is the frequency of a pattern p_i in tc, and $f(p, tc)$ is the total number of patterns present in tc.

$$PF(p_i, tc) = log\left(1 + \frac{f(p_i, tc)}{f(p, tc)}\right) \quad (10)$$

The inverse document frequency IDF for a pattern p_i is computed to provide the relevance of pattern p_i with the training corpus using equation 11, where $n(X)$ is the total number of input tweets in the rumor training corpus tc and $f(x, p_i)$ represents the total number of input tweets in which pattern p_i occurs.

$$IDF_{p_i} = log\left(1 + \frac{n(X) \in tc}{f(x, p_i)}\right) \quad (11)$$

Thus, the pattern frequency-inverse document frequency $PF - IDF$ is calculated by scalar multiplication of equations 10 & 11, as shown in equation 12, and it is considered as the final weight score for a pattern p_i.

$$PF-IDF(p_i,tc) = PF_{(p_i,tc)} \times IDF_{p_i} \qquad (12)$$

We arrange the patterns in each category of the list in descending order of their weight score for selecting top-k check-worthy patterns. Finally, the top-k patterns from all three lists are combined to make a hybrid patterns list that is useful in detecting the rumor.

3.5.1 Semantic Vectorization

We have applied embedding on the input tweets to preserve the semantic similarity with the patterns. For each input tweet X_i of the training corpus, we split it into n-grams and generate a semantic vector of each n-gram using word embedding. For example, if a tweet has m number of n-grams, then the semantic vector of the tweet is represented as a matrix of $\big[[|n| \times [d] \quad |n| \times [d] \quad \dots \quad |n| \times [d]\big]_{1\times m}$, where $|n|$ is number of words in n-grams and d is the dimension of the word embedding vector.

The vectorization step is also applied over the identified patterns. It enriches patterns and provides semantic information by using word embedding. For preserving the semantic relations, we make a semantic vector for a pattern using the same word embedding. The semantic vector of a pattern is a matrix of order $|v_i| \times [d]$, where $|v_i|$ represents the number of words in a pattern p_i.

3.6 Feature Vector Generation

In this step, tweets are converted into a feature vector using Cosine similarity. The semantic vector of both the patterns and n-grams of the input tweets are matched, and the mean similarity score is calculated. For example, a list consists of top-k patterns; for an i^{th} pattern p_i, the similarity score is obtained by matching it with each n-gram. Furthermore, we calculate the mean of the similarity score. This step repeats until all the patterns from the list are matched with n-grams of the input tweets. Since there are total k patterns, the size of the final feature vector is k. For a pattern p_i, the numeric value for the feature vector is denoted by χ_{p_i} and calculated using equation 13, where m is the total number of n-grams in input tweet X_i and $Cosine(p_i, X_{i_m})$ is a similarity score of a pattern p_i and a n-gram of input tweet X_i.

$$\chi_{p_i} = \frac{\sum_1^m Cosine(p_i, X_{i_m})}{m} \qquad (13)$$

4 Experimental Setup and Results

4.1 Dataset

We conduct our experiment on a publicly available PHEME dataset described in (Zubiaga et al., 2017). The dataset contains a collection of tweets with their reactions (direct or nested) and metadata of five breaking news events. There is a total of 5802 source tweets in which 1972 are labeled as a rumor and 3830 as a non-rumor. This dataset contains less number of rumors in comparison to non-rumors. Therefore, another variant of a balanced dataset is created from the original dataset, where both rumor and non-rumor have an equal number of 1972 instances. We conduct our experiment on both variants of the datasets. The detailed statistics of the dataset is presented in Table 1.

Table 1: Statistics of the dataset

Events Name	Source Tweets	Reactions	Rumors	Non-Rumors	Total
Charlie Hebdo	2079	36189	458	1621	38268
Sydney Siege	1221	22775	522	699	23996
Ferguson	1143	23032	284	859	24175
Ottawa Shooting	890	11394	470	420	12284
Germanwings Crash	469	4020	238	231	4489
Total	5802	97410	1972	3830	103212

4.2 Evaluation Metrics

This section discusses the formal description of performance evaluation metrics for classification. The three standard data mining evaluation metrics- *Precision, Recall*, and *F1-Score* are defined in the equation 14, 15, and 16, respectively. For defining the evaluation metrics, we have used the concept of *True-Positive (TP)*, i.e., the number of rumorous tweets identified correctly, *False-Positive (FP)*, i.e., the number of non-rumor tweets identified as rumorous, and *False-Negative (FN)*, i.e., the number of rumorous tweets identified as non-rumor. *Precision* evaluates the correctness of the classifier, and *Recall* evaluates the completeness of coverage of the classifier, while *F1-Score* provides the way to combine the contribution of both *precision* and *recall* evenly by using harmonic mean.

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

4.3 Evaluation Results and Comparative Analysis

The experimental evaluation is performed using four machine learning classification algorithms *support vector machine (SVM), gradient boosting (GB), conditional random field (CRF)*, and *logistic regression (LR)* over both the variants of benchmark dataset. The proportion of train and test sets in the proposed approach is $8:2$. We have implemented the above classification models using `scikit-learn python` library.

Following pre-processing, the unigram frequency matrix of graph words is generated. Those unigrams with a frequency of less than ten are discarded. After that, we applied *clustering coefficient* and *eigenvector centrality* on the graph to extract the two prevalent categories of words. The graph generation and its operations are implemented using the `networkx` package for the `Python` programming language. We have considered different threshold values to extract top words from both categories. Words of these two categories are located in the graph and their immediate connections i.e., $e_{ij} \in E$ of a node v_i with neighbors N_i, where $v_j \in N_i$, are retrieved. We used the sliding window of size 2; therefore, patterns of three words are created by combining a node with its neighbor in a way that not all three words are from the same category.

To make patterns generic, we incorporated negative emotional lexicons and skeptical words. As a result, we curated three lists of patterns named rumor patterns, emotional patterns, and skeptical patterns. The word clouds in figure 2 present words of the top-100 patterns for each category in which the font size of a word is directly proportional to its relevance score. It can be observed from these word clouds that the words with larger font sizes are significantly related and explain the context of the underlying categories. We ranked the patterns through *tf-idf* and selected the top-k patterns empirically. We have used 100-dimensional word-level pre-trained `GloVe` embedding trained over the `Twitter` dataset with 27 billion tokens. Those words that do not exist in the pre-trained word vectors are ignored. We split each input tweet into trigrams since each pattern has three words. While splitting the input tweets, 2 rumors and 15 non-rumor are ignored since they have less than three words, and trigrams can not be created. The patterns and trigrams of tweets are represented in the 3×100 dimensions; we have taken their centroid value and reduced it to the 1×100 dimensions. At last, the mean similarity value of a pattern with each trigram of tweets is calculated. The top-k patterns produce a k-dimensional feature vector.

The propose approach is compared with the following baseline methods and state-of-the-art approaches.

Baseline 1: In this method, only the top-k rumor patterns are incorporated to generate the feature vector.

Baseline 2: In this method, the top-k patterns from rumor as well as skeptical patterns are incorporated to generate the feature vector.

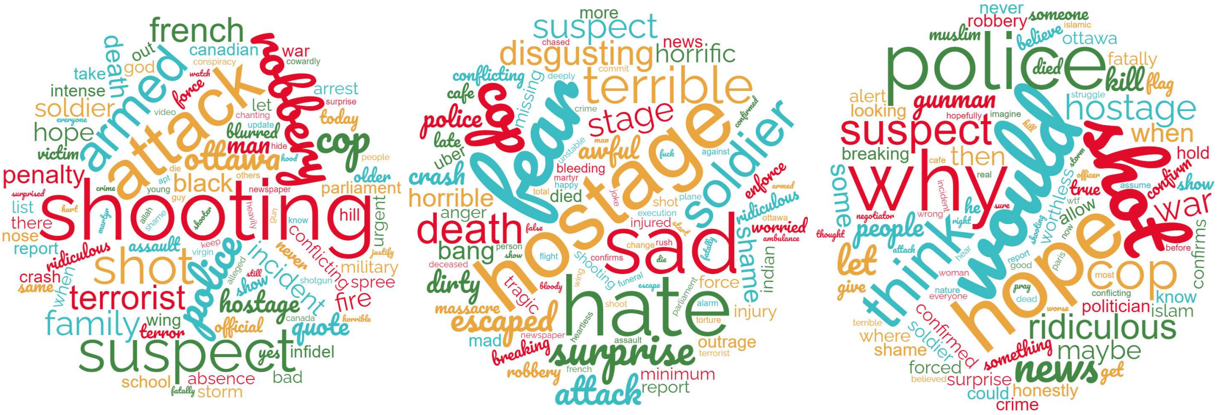
Baseline 3: In this method, the top-k patterns from rumor as well as emotional patterns are incorporated to generate the feature vector.

Ajao et al. (2019): This approach considered the relationship of rumors with the sentiments of the social media post. It has used emotional words to detect the sentiment-aware misinformation.

Abulaish et al. (2019): This approach used the graph-based approach for rumor detection that incorporated the sentimental aspects, such as anxiety and doubtful terms from the social media post.

Zubiaga et al. (2017): This approach has learned through the sequential dynamics of the social media post. The content features and user social features have experimented with *conditional random fields (CRF).*

We have compared the proposed approach with three baselines and three state-of-the-art approaches.



(a) Rumor patterns

(b) Emotional patterns

(c) Skeptical patterns

Figure 2: A word-cloud representing words of the top-100 patterns for each category

Table 2: Comparative performance evaluation results of our proposed approach with state-of-the-art approaches and baseline methods over the original dataset

Approach	GB			SVM			CRF			LR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline1	82.48	72.44	77.13	82.07	62.61	71.03	79.81	65.68	72.04	76.09	52.35	62.03
Baseline2	79.38	76.35	77.83	81.20	63.32	71.15	82.32	72.34	77.00	80.00	53.71	64.27
Baseline3	85.80	87.82	86.80	87.53	91.45	89.45	86.34	82.56	84.40	83.84	76.50	80.00
Ajao et al. (2019)	84.80	85.12	84.96	86.68	85.82	86.24	86.43	84.64	85.52	83.83	85.11	84.46
Abulaish et al. (2019)	56.26	55.40	55.82	41.30	45.56	43.32	64.62	60.10	62.28	40.80	41.93	41.11
Zubiaga et al. (2017)	52.43	54.19	53.29	36.60	44.78	40.27	69.19	54.59	61.02	33.10	40.72	36.51
Proposed Approach	92.83	94.02	93.42	93.48	91.88	92.67	91.06	88.97	90.00	90.97	86.11	88.47

Table 2 summarizes the comparative results in terms of *precision*, *recall*, and *f1-score* over the original dataset. It can be observed that the proposed approach outperforms all other approaches for all four classification algorithms. Our best result is obtained through the *gradient boosting*. It can also be observed that *gradient boosting* achieved highest *recall* and *f1-score*, whereas *SVM* achieved highest *precision*. To assess the effect of class imbalance, we repeated the same set of experiments with the variant of a balanced dataset. As shown in table 3, *gradient boosting* scored the highest value of *precision* and *f1-score*, whereas *CRF* achieved the highest *recall* value. It can also be observed that the *recall* and *f1-score* values are better over the balanced dataset for all four classification algorithms.

Figures 3 and 4 present a visualization of the comparative analysis results of the proposed approach with three state-of-the-art techniques over both original and balanced datasets, respectively. It can be observed that the proposed approach performs significantly better than all state-of-the-art approaches. The improvements of the proposed approach over the best state-of-the-art approach range from $4.01 - 8.46\%$ for the original distribution of the dataset and $3.93 - 6.33\%$ for the balanced dataset. The out-performance consistency of our proposed approach is maintained for both variants of the dataset with a difference of $1.03 - 3.85\%$ in the *f1-score* value. In contrast, the state-of-the-art approaches are inconsistent; some methods improved 12.07% in the *f1-score* value, whereas the performance of a few decreases. The reason for the consistent performance of the proposed approach is that the graph-based approach covers the topical words and the writing style of the social media post, whereas the word embed-

Table 3: Comparative performance evaluation results of our proposed approach with state-of-the-art approaches and baseline methods over the balanced dataset

Approach	GB			SVM			CRF			LR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline1	82.49	79.54	80.99	79.62	79.96	79.79	81.21	82.38	81.79	76.82	75.53	76.17
Baseline2	82.77	85.86	84.29	76.35	85.24	80.55	80.54	87.24	83.75	76.57	78.79	77.66
Baseline3	91.10	92.83	91.95	89.90	95.10	92.42	88.13	95.04	91.45	85.80	94.30	89.85
Ajao et al. (2019)	86.88	89.21	88.03	88.00	89.56	88.77	84.50	83.86	84.18	85.11	82.37	83.71
Abulaish et al. (2019)	61.22	59.46	60.32	52.41	50.86	51.62	82.04	67.98	74.35	63.56	60.02	61.74
Zubiaga et al. (2017)	58.63	60.40	59.50	51.13	50.75	50.94	80.00	62.53	70.19	52.43	45.66	48.81
Proposed Approach	95.36	95.36	95.36	91.72	95.78	93.70	91.18	95.88	93.47	90.82	93.88	92.32

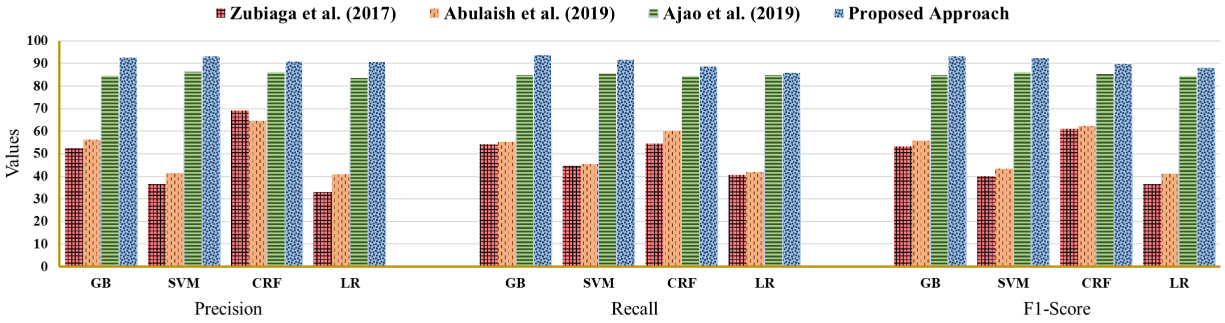


Figure 3: Comparative performance evaluation analysis over the original dataset

ding covers the higher dimensional semantic space.

5 Conclusion and Future Work

In this paper, we have presented a rumor detection framework that uses a graph-based approach to leverage tweets and reactions for extracting rumorous patterns. A graph-based learning representation is used to capture contextual information. The inquisitive, skeptical, sentimental, and emotional natures of social media users are identified through their writing styles in the reactions to a tweet. The semantic relations are preserved through semantic vectorization, based on word embedding. The hybrid top-k patterns are extracted from all three categories – rumor, emotional, and skeptical that are used to train the rumor detection model. The experimental results explained that our proposed approach significantly improves the rumor detection task and outperforms the state-of-the-art methods. As a result, it is concluded that utilizing emotional and skeptical words makes the detection system more effective. It can also be said that the pattern-based approaches result in more representative and smaller size models than the word-based approaches. The patterns also retain the syntactic sense. In this study, we have considered three-words patterns because long patterns are infrequent. The proposed work can be extended to improve the pattern ranking mechanism to maximize the coverage of patterns.

References

Muhammad Abulaish, Nikita Kumari, Mohd Fazil, and Basanta Singh. 2019. A graph-theoretic embedding-based approach for rumor detection in twitter. In *Proceedings of the 18th IEEE/WIC/ACM International Conference on Web Intelligence, Thessaloniki, Greece*. Association for Computing Machinery, 466–470.

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *Proceedings of the 44th IEEE International Conference on Acous-*

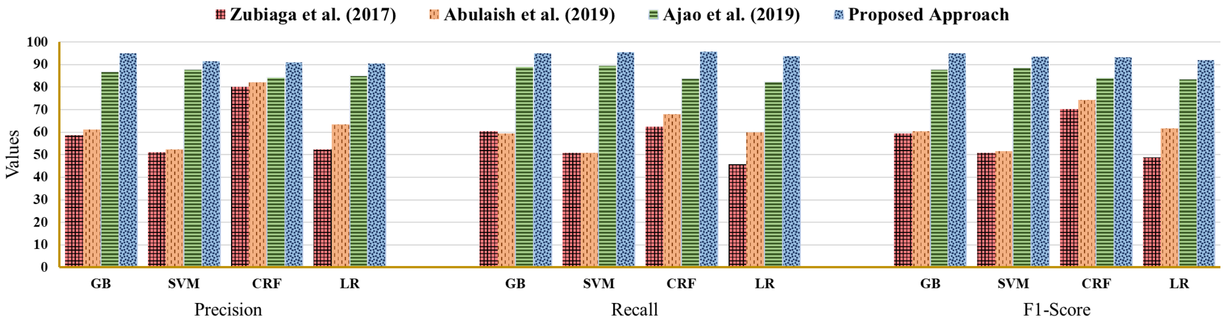


Figure 4: Comparative performance evaluation analysis over the balanced dataset

tics, Speech and Signal Processing (ICASSP), Brighton, UK. IEEE, 2507–2511.

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2021. Rumour Detection Based on Graph Convolutional Neural Net. *IEEE Access* 9 (2021), 21686–21693.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India*. Association for Computing Machinery, 675–684.

Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and Guohui Li. 2019. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China*. Association for Computing Machinery, 569–578.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada*. Association for Computational Linguistics, 475–480.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *Proceedings of 28th The World Wide Web Conference, San Francisco, CA, USA*. Association for Computing Machinery, 3049–3055.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, California, USA*. Association for Computational Linguistics, 31–41.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland*. Association for Computing Machinery, 1165–1174.

Nir Rosenfeld, Aron Szanto, and David C Parkes. 2020. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of 29th The Web Conference, Taipei Taiwan*. Association for Computing Machinery, 1018–1028.

Kefei Tu, Chen Chen, Chunyan Hou, Jing Yuan, Jundong Li, and Xiaojie Yuan. 2021. Rumor2vec: a rumor detection framework with joint text and propagation structure representation learning. *Information Sciences* 560 (2021), 137–151.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of* *the 24th International Conference on World Wide Web, Florence Italy*. Association for Computing Machinery, 1395–1405.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. *Advances in Knowledge Discovery and Data Mining, Springer* 12085 (2020), 354.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–36.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Proceedings of International Conference on Social Informatics, Oxford, UK*. Springer, 109–123.

Siamese AraBERT-LSTM Model based Approach for Arabic Paraphrase Detection

Adnen Mahmoud[1,2] **Mounir Zrigui**[1]
[1]University of Monastir, Research Laboratory in Algebra, Numbers Theory and Intelligent Systems RLANTIS, Monastir 5000, Tunisia
[2]University of Sousse, Higher Institute of Computer Science and Communication Techniques ISITCom, Hammam Sousse 4011, Tunisia
mahmoud.adnen@gmail.com mounir.zrigui@fsm.rnu.tn

Abstract

Paraphrase detection allows identifying the degree of likelihood between source and suspect sentences. It is a critical machine learning problem in computational linguistics. This is due to the expression variability and ambiguities especially in Arabic language. Previous neural models have yielded promising results, but are computationally expensive. They cannot directly align long-form sentences expressing different meanings. To address this issue, Siamese neural network is proposed for Arabic paraphrase detection based on deep contextual semantic textual similarity. Despite that the pre-trained word embedding models have advanced NLP, they ignored the contextual information and meaning within the sentence. In this paper, the potential of deep contextualized word representations was firstly investigated using Arabic Bidirectional Encoder Representation from Transformers (AraBERT) as an embedding layer. Then, Long Short Term Memory (LSTM) modeled high-level semantic knowledge. Finally, cosine distance identified the degree of semantic textual similarity. Using our own generated corpus, experiments showed that the proposed model outperformed state-of-the-art methods, in terms of F1 score.

1 Introduction

The accumulation of textual data exchanged on the web over time has increased the potential source of paraphrase (Karaoglan et al., 2016). Its identification has become increasingly challenging especially in the case of Arabic language due to its richness of ambiguous specificities (Sghaier and Zrigui, 2020). The same sentence can be reformulated in different ways using semantically similar words (Mahmoud and Zrigui, 2021a). This task allows modelling and identifying the semantic interactions between sentence pairs. It represents a challenge in the area of information retrieval and Natural Language Processing (NLP). Recently, several neural models for word embedding have been introduced like word2vec (Mikolov et al., 2014) or GloVe (Pennington et al., 2014). The produced vectors representations of sentence pair are used as inputs to measure the similarity between them. However, these models provided fixed representation for each word and did not capture its context in different sentences (Mahmoud and Zrigui, 2019a).

To deal with this drawback, contextualized word representation methods such as ELMo (Peters et al., 2018) and BERT (Babi et al., 2020) have become prevalent and received a lot of attention for obtaining sentence representations. They learnt efficiently the contextualized word representations from deep bidirectional language model pre-trained on large text corpora or via utilizing the encoder of transformer (Babi et al., 2020). Better sentence representations are captured from each generated word embedding based on its surrounding context. Therefore, we are motivated to use them for sentence modelling and Arabic paraphrase detection.

In this paper, the main objectives are focused on studying how the applications of Bidirectional Encoder Representations from Transformers (BERT) and neural networks models are suitable for semantic equivalence assessment and Arabic paraphrase detection. Indeed, the contextual features are firstly extracted using Arabic BERT (AraBERT) model. Then, the resulted embedded vectors are trained by applying deep Siamese Long Short Term Memory (LSTM) model for sequential data modelling. Finally, semantic similarity scores are identified. To conduct experiments, paraphrased corpus is proposed preserving semantic and syntactic properties of original sentences. This paper is organized as follows: First, state of the art on paraphrase detection is presented in section 2. Then, the proposed approach is detailed in section 3. Subsequently, the experiments are described in section 4. Finally, we end by a conclusion and future work in section 5.

2 State of the Art

Although processing language and comprehending the contextual meaning is an extremely complex task, paraphrase detection is a sensitive field of research for specific language (Mahmoud and Zrigui, 2019b). Following the literature, numerous neural models were proposed for modelling semantic similarity among sentence pair. They have gained promising results in major NLP tasks distinguish supervised and unsupervised approaches.

Unsupervised methods use pre-trained word/phrase embeddings directly for the similarity task without training a neural network model on them which supervised ones do (Aliane and Aliane, 2020).

Word2vec is one of the most popular unsupervised methods. It generates words embeddings according to their semantics in the sentence. Some researches were already used it to detect similarity between texts. Veisi et al. (2022) employed word2vec and cosine distance for Persian text similarity detection. The same approach was proposed by Gharavi et al. (2016). Indeed, the generated word vectors from word2vec model were averaged to produce sentence vector. Then, the Jaccard coefficient was used to report plagiarism cases. For Arabic language, Nagoudi et al. (2018) detected verbatim and complex reproductions using fingerprinting and word embedding. Next, different NLP techniques were combined like word alignments, Inverse Document Frequencies (IDF) and Part-Of-Speech (POS) weighting to identify the most descriptive words in each textual unit. Similarly, Yalcin et al. (2022) indexed each document according to the grammatical classes of their n-grams. The aim was to access rapidly to sentences that were possible plagiarism candidates. Then, word2vec and Longest Common Subsequence (LCS) were combined for measuring semantic similarity.

Unlike word2vec, GloVe model does not rely on local context information, but incorporates global statistics through word co-occurrences to obtain word vectors (Mahmoud and Zrigui, 2021b). For English text similarity identification, the superiority of GloVe compared to word2vec was demonstrated by Hindocha et al. (2019) and Mohammed et al. (2019).

Other works adopted FastText model. For example, Iqbal et al. (2021) investigated several word embedding techniques (word2vec, GloVe, and FastText) for Bengali semantic similarity. Experiments showed that FastText with cosine distance were the most suitable for this task.

Recently, context-depending embedding techniques have been introduced. They used transformers and long and short-term memory techniques to convert a word into n-dimensional vector. To enhance unsupervised sentence similarity methods, Ranashinghe et al. (2019) enhanced various context-based models: Embedding from Language Models (ELMO) (Peters et al., 2018), Bidirectional Encoder Representations from Transformers BERT (Babi et al., 2020), Flair (Akbik et al., 2019) and stacked embedding on different datasets: English, Spanish and Biomedical. The best experimental results were obtained with stacked embeddings of ELMO and BERT.

On the *supervised methods* side, deep pairwise fine-grained similarity network is based on Siamese architecture. Two identical neural networks sharing the same weights. The resulted output vectors are fed to a join function for paraphrase prediction (Mahmoud and Zrigui, 2021c).

Convolutional Neural Networks (CNN) were efficient to extract the most descriptive n-grams of different semantics through convolution and pooling layers. This model was applied for sentence modeling and semantic text similarity computation as shown by Shao (2017) and He and Lin (2015) for English, and Mahmoud and Zrigui (2017) for Arabic. Recurrent neural networks have recently shown promising results for analysing sequential data and modelling long-term dependencies within sentence (Haffar et al., 2021). For modelling context and structure of sentence, Tai et al. (2015) proposed tree-structured LSTM model while Liu et al. (2019) introduced a Multi-Layer Bidirectional LSTM (Bi-LSTM) and TreeLSTM model. As shown in (Hambi and Benabbou, 2019), LSTM model learnt from the output of doc2vec while CNN model learnt thereafter the most relevant features from documents for plagiarism detection. As described in (Othman et al., 2021), LSTM model was augmented with an attention mechanism to extract the most representative words within questions. Then, CNN retrieved relevant questions. Hamza et al. (2020) extracted semantic and syntactic relations between words using ELMo model. Then, different deep neural models were analyzed such as simple models, CNN and RNN (Bi-LSTM, GRU) mergers models, and ensemble models. Then, a concatenation operation was used as a merge operation and Softmax function for Arabic similar questions classification. Recently, Meshram and Kumar (2021) demonstrated that deep contextualized word representations using BERT model became a better way for feature extraction from sentences. LSTM was thereafter applied for high level features knowledge. Then, Manhattan distance was used for similarity identification.

Following the literature, our approach combines the power of unsupervised learning through contextualized word embedding and supervised deep pairwise similarity network that outperformed state-of-the-art. Our motivation is to use them for sentence modelling and paraphrase identification in Arabic. This is due to the rich language of features that made its processing difficult than other languages (Meddeb et al. 2021). It is non-vocalized, non-concatenative, homographic, agglutinative and derivational, which needs deep understanding of textual components (Haffar et al., 2020)

3 Proposed Approach

In this section, deep contextual embedding based similarity approach is presented briefly for Arabic paraphrase detection. It is based on a Siamese neural architecture. It has proven its relevance for learning sentence semantic comparability through two identical sub-networks that are fit for handling sentence pair and likeness measure. The phases constituting the proposed architecture are described in Fig. 1:

1) Firstly, text features are extracted using the Arabic Bidirectional Encoder Representations from Transformers (AraBERT) process.
2) The embedded vectors are subsequently trained by using Long Short Term Memory (LSTM) recurrent neural network model.
3) Finally, similarity scores are determined for each sentence pair, and the semantic textual similarity is learned.

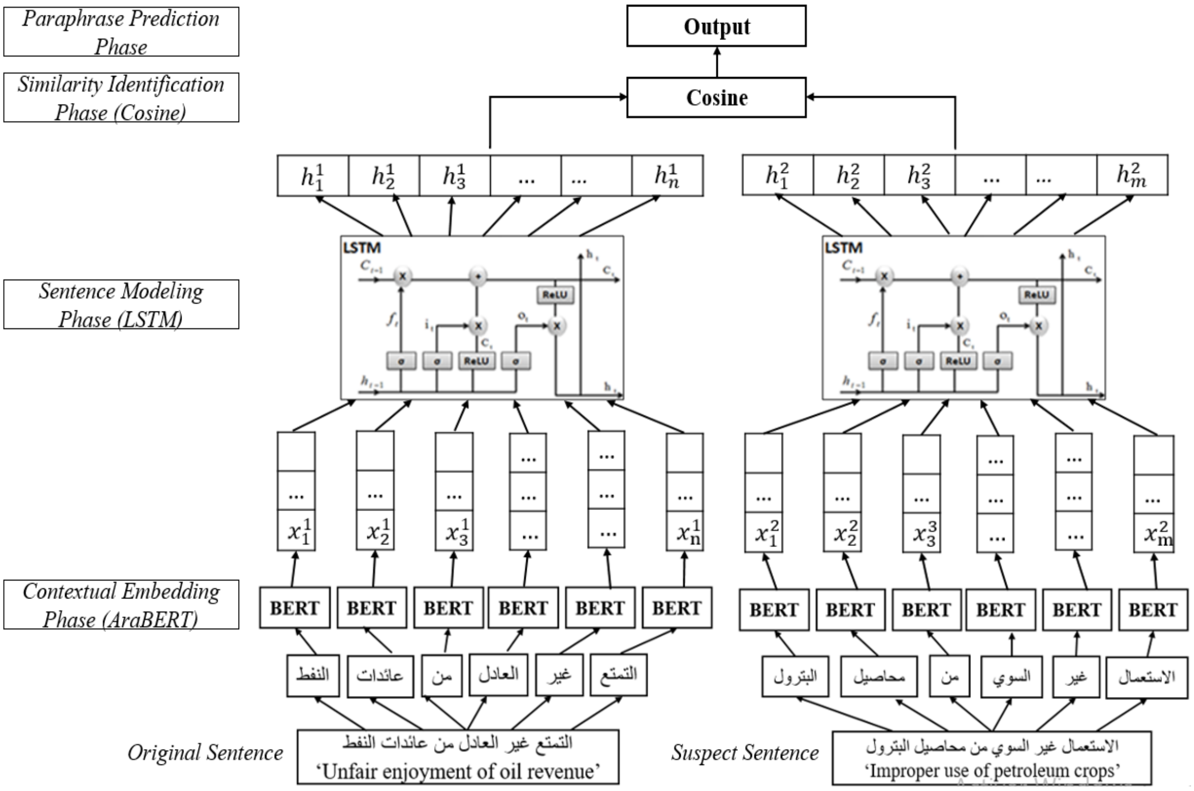

Paraphrase Prediction Phase
Similarity Identification Phase (Cosine)
Sentence Modeling Phase (LSTM)
Contextual Embedding Phase (AraBERT)
Output
Cosine
h_1^1 h_2^1 h_3^1 ... ... h_n^1
h_1^2 h_2^2 h_3^2 ... ... h_m^2
LSTM
ReLU
x_1^1 x_2^1 x_3^1 x_n^1
x_1^2 x_2^2 x_3^3 x_m^2
BERT
النفط عائدات من العادل غير التمتع
البترول محاصيل من السوي غير الاستعمال
Original Sentence
التمتع غير العادل من عائدات النفط
'Unfair enjoyment of oil revenue'
Suspect Sentence
الاستعمال غير السوي من محاصيل البترول
'Improper use of petroleum crops'

Fig. 1: Proposed architecture.

3.1 Arabic BERT (AraBERT) Embedding Phase

The use of word embeddings allows to effectively detect the syntactic and semantic similarities between words. In this work, we resorted to the AraBERT model which outperformed previous language models like word2vec, GloVe, etc. It addressed effectively ambiguity in which multiple vector representations could be extracted for the same word based on its context. More precisely, AraBERT is a bidirectional transformer for generating sentence representation by learning the context of each word on all of its surroundings in the sentence. Because AraBERT uses sub-words as a unit instead of words, the source S and target T sentences are tokenized into words. Then, the obtained tokens are thereafter encoded by the AraBERT model. For the input sequence of N tokens $\{w_1, \ldots, w_n\}$, we obtain the final hidden states $\{h_1, \ldots, h_n\ \}$representing the output of the transformer as denoted in Eq. (1) :

$$h_i = AraBERT(w_1, \ldots, w_n) \qquad (1)$$

To generate the representations of source h_s and target h_t sentences, we use the mean_pooling process applied on the outputs of AraBERT model. It consists computing the means over each vector dimension as follows in Eq. (2) and Eq. (3):

$$h_s = mean\_pooling(h_1, \ldots, h_m) \qquad (2)$$

$$h_t = mean\_pooling(h'_1, \ldots, h'_n) \qquad (3)$$

Where: m and n are the lengths of source and target sentences.

3.2 Sentence Modeling Phase

The major reason for relying on Long-Short Term Memory (LSTM) recurrent neural network is its proven performance to capture short and long-term dependencies, model variable-length of sequential data and prevent the vanishing gradient problem of RNN. LSTM is characterized by a memory cell that is capable of maintaining its state over time, and internal mechanisms called gates to regulate the information flow.

Given the input vector x_t, hidden sate h_t and memory state c_t, the updates in LSTM are performed as denoted in Eqs. (4, 5, 6, 7, 8 and 9):

$$i_t = ReLU(W_i.x_t + U_i h_{t-1} + b_i) \qquad (4)$$

$$f_t = ReLU(W_f.x_t + U_f h_{t-1} + b_f) \qquad (5)$$

$$\check{c}_t = ReLU(W_c.x_t + U_c h_{t-1} + b_c) \qquad (6)$$

$$c_t = i_t \odot \check{c}_t + f_t \odot c_{t-1} \qquad (7)$$

$$o_t = ReLU(W_o.x_t + U_o h_{t-1} + b_o) \qquad (8)$$

$$h_t = o_t \odot Tanh(c_t) \qquad (9)$$

Where : i_t, f_t, c_t, o_t are input, forget, memory and output gates at time t ; $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o$ are weight matrices ; b_i, b_f, b_c, b_o are the bias vectors ; ReLU (Rectified Linear Unit) is the activation function and $\odot$ denotes the Hadamart product of matrices.

3.3 Similarity Identification Phase

Once we have the vectors (A, B) that capture the underlying meaning of sentence pair, the semantic similarity is computed using Cosine similarity measure as defined in Eq. (10):

$$Sim(A,B) = \frac{A.B}{||A||.||B||} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \qquad (10)$$

Where: A_i and B_i are the components of the vectors A and B, respectively.

For prediction, the obtained scores of $Sim(A, B)$ are converted into probabilities $P \in [0,1]$ as defined in Eq. (11). To decide whether or not $A\ and\ B$ are paraphrased (i.e., semantically equivalent), the obtained P is compared to a threshold $\alpha = 0.25$:

$$P = \frac{Sim(A,B)}{10} = \frac{Sim(A,B) \times 5}{100} + \frac{50}{100} \qquad (11)$$

4 Experiments and Discussion

4.1 Dataset

To tackle the lack of publicly available paraphrased corpora in Arabic, we intend to develop our own corpus. The main idea is to capture more semantic features from sentence pairs. This is done by combining word2vec model and POS weighting for better-paraphrased sentence generation. It consists of the following operations:

Data collection. The Open Source Arabic Corpora (OSAC) is used as a source corpus from which passages of texts are extracted and replaced semantically. To do this, vocabulary model is proposed from which original words are replaced. It is collected from various resources (i.e. Arabic Corpora Resource (AraCorpus), King Saud University Corpus of Classical Arabic (KSUCCA) and a set of Arabic papers from Wikipedia) including more than 2.3 billion words.

Data preprocessing. To remove worthless data and reduce thereafter the time required for further processing, preprocessing operations are applied:

(1) Unnecessary data (e.g. extra white spaces, titles numeration, non-Arabic words) are removed.

(2) Some writing forms are normalized (e.g., *Hamza* "أ" and *Taa Marboutah* "ة" to "ا" and "ه".)).

(3) Sentences are tokenized into words to reduce lexical parsimony.

Paraphrased corpus generation. Semi-artificial approach is proposed as follows:

(1) Given a random variable $a \leq x \leq b$, the degree of paraphrase D is defined by applying a random uniform function, as denoted in Eq. (12):

$$f(x) = \begin{cases} \frac{1}{b-a}, a \leq x \leq b \text{ and } (b-a) \in [1.33, \ldots, 2.22] \\ 0, \text{ otherwise} \end{cases} \tag{12}$$

(2) The number of words to replace P is defined from the OSAC source corpus of size N. It is defined according to $D \in [45\%, \ldots, 75\%]$, as follows in Eq. (13):

$$P = N \times D \tag{13}$$

(3) To replace source words according to an index chosen randomly, random shuffle function is used.

(4) Since paraphrase allows replacing the meaning of original words semantically, synonyms are identified from the vocabulary using Skip gram model. It predicts the context of middle word according to the unique word representation in a surrounding window as input. The original word is replaced by its most similar one that has the same grammatical class from the created vocabulary. In this way, original and paraphrased sentences will have the same syntactic changes with similar words semantically. The combination of word2vec and POS is reported to be good in capturing syntactic and semantic features of words.

(5) Human judgments are used for final corpus validation.

Table 1 summarizes some obfuscations forms created through our proposed approach:

Original sentence	التمتع غير العادل من عائدات النفط. 'Unfair enjoyment of oil revenue.'	
w_i	Obfuscation forms	
	Synonym substitution	Add / remove of words
التمتع *'enjoyment'*	الاستعمال 'use'	الاستعمال 'use'
غير *'not'*	غير 'not'	-
العادل *'fair'*	السوي 'proper'	الخاطئ 'wrong'
من *'of'*	من 'of'	ل 'of'
عائدات *'revenue'*	محاصيل 'crops'	محاصيل 'crops'
النفط *'oil'*	البترول 'petroleum'	البترول 'petroleum'
Suspect sentences	الاستعمال غير السوي من محاصيل البترول 'Improper use of petroleum crops'	الاستعمال الخاطئ لمحاصيل البترول 'Misuse of petroleum crops'

Table 2: Examples of Paraphrased sentences in Arabic

Table 2 describes the data used for training and testing the proposed approach:

Corpora	Models	Total pairs	Paraphrased pairs	Original pairs
OSAC	Train	3,600	2,400	1,200
	Test	1,500	1,000	500
Semeval	Test	250	196	94

Table 2: Experimental dataset

4.2 Neural Models Configuration

Word2vec. Table 3 illustrates the parameters of word2vec algorithm that are useful to capture the right synonyms of the target word and increase the quality of Arabic paraphrased sentence generation:

Parameters	Values
Vocabulary size	More than 2.3 billion words
Vector dimension	300
Window size	3
Minimum count	≤ 5
Workers	8
Epochs	7

Table 3: Parameters of word2vec model

AraBERT[1]**.** This model is employed for sentence modelling. As shown in Table 4, the parameters (i.e., number of epochs, hidden layers, attention heads, and hidden size) are fixed according to the dataset and the memory reserved. The overall model is trained by Adam optimizer.

Parameters	Values
Hidden layers	4
Attention heads	4
Hidden size	256
Dropout	0.1
Optimizer	Adam
Epochs	50
Activation function	ReLU

Table 4: Parameters of AraBERT model

LSTM. For sentence modeling, the parameters that increased the performance of the proposed LSTM model are summarized in Table 5:

Parameters	Values
Hidden units number	256
Activation function	ReLU
Loss probability	0.2
Optimizer	Adam
Batch size	100

Table 5: Parameters of LSTM model

4.3 Evaluation Metrics

F1 score is a measure of the proposed model's accuracy on the generated dataset. It is defined as the harmonic mean of the precision and recall ranging in [0, 1], as defined in Eq. (14) (Mahmoud and Zrigui, 2021d):

$$2 \times \frac{P \times R}{P + R} \qquad (14)$$

Where: precision P is the fraction of true positive examples among the ones that the model classified as positive; recall R is the fraction of examples classified as positive among the total number of positive examples.

4.4 Discussion

The performances of our approach using the proposed corpus and SemEval dataset are comparable to detect semantic meaning of words, which also depends on deep contextual embedding. Experimental results are shown in Table 6:

Corpora	Models	F1
OSAC	GloVe-Cosine	0.7750
	AraBERT-Cosine	0.8050
	GloVe-LSTM	0.8731
	AraBERT-LSTM	0.8975
SemEval	GloVe-Cosine	0.7650
	AraBERT-Cosine	0.7850
	GloVe-LSTM	0.8397
	AraBERT-LSTM	0.8650

Table 6: Experimental results

Experiments demonstrated that GloVe model with cosine similarity achieved the lowest F1 score (77.5% with OSAC and 76.5% with SemEval). This is due to the fact that GloVe worked on word level and cannot cope with Arabic language morphology. Indeed, different senses of the words are combined into one vector which can result a confused representation of the ambiguous Arabic language. As a solution, we proposed to integrate it topped with LSTM layer. It increased the number of learnable weights and paraphrase prediction layer. It captured efficiently sentence semantics better. As expected, adding a pairwise sentence similarity sub-networks improved the performance of our model achieving the best F1 scores: 87.31% using OSAC and 83.97% using SemEval.

Instead of using GloVe model, the impact of contextualized word embedding AraBERT is examined. It consistently improved the results in all the proposed models. While Arabic language is highly derivational language mutating many morphological variations of each word, AraBERT model was able to generate vector representation of any arbitrary words and was not limited to the vocabulary space. It obtained the highest F1 scores: 89.75% with OSAC and 86.50% with SemEval.

Furthermore, we can notice that low performances are obtained when using SemEval corpus compared to the proposed OSAC corpus. Based on the complex nature and structure of Arabic language, this limit is related to the nature of the corpus, its language and the proposed paraphrase detection approach: Compared to the OSAC corpus, SemEval corpus has a relatively small vocabulary and training examples. That's why, the performance of AraBERT based approach hasn't been affected. We can deduce

[1] https://github.com/aub-mind/arabert/tree/master/arabert

also the effectiveness of the model when facing small amount of data and more generally its suitability when dealing with low resources languages like Arabic.

As depicted in Table 7 and Fig. 2, final experimental results are competitive with those obtained with state-of-the-art methods achieving the best F1 score with AraBERT-LSTM model. It was efficient for text similarity prediction and paraphrase detection:

References	Corpora	Models	F1
Ayata et al. (2017)	SemEval English	Word2vec-LSTM	0.587
Peinelt et al. (2020)	SemEval English	tBERT-LDA	0.524
	MSRP English	tBERT-LDA	0.884
Meshram and Kumar (2021)	SICK, STS, clinical dataset	Word2vec-BERT	0.889
		GloVe-BERT	0.864
		BERT-BERT	0.896
Ours	OSAC Arabic	AraBERT-LSTM	0.897
	SemEval Arabic	AraBERT-LSTM	0.865

Table 7: Comparison with state-of-the-art methods

For sentiment analysis and Tweets similarity, word2vec algorithm was useful for features representation while LSTM model was thereafter efficient for avoiding the long-term dependency problem, as shown by Ayata et al. (2017). Recently, Meshram and Kumar (2021) and Peinelt et al. (2020) approved that the use of BERT model as a contextualized word embedding was even applying it on topics for English text similarity. Compared to the state-of-the-methods, the overall best performing method was obtained using AraBERT for contextualized word embedding in Arabic, LSTM for sentence pair modelling and cosine for similarity prediction as demonstrated in our proposed final approach. It achieved the highest F1 scores (89.75% using OSAC and 86.50% using SemEval).

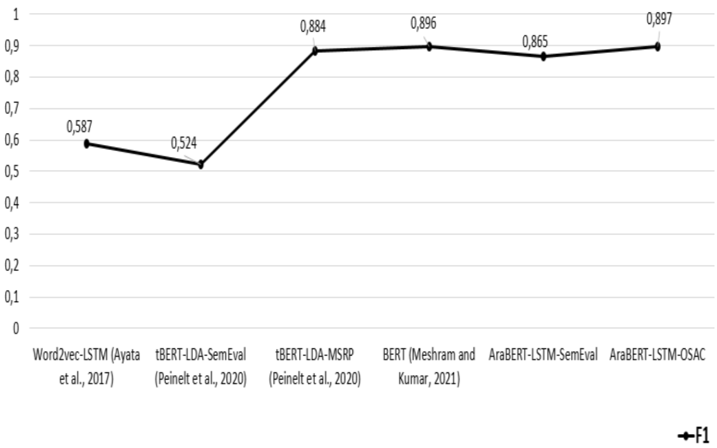


Fig. 2: Comparison with state-of-the-art methods

5 Conclusion and Future Work

In this paper, we introduced an Arabic paraphrased corpus preserving semantic and syntactic features of sentences. Original words are replaced by their most similar ones that had the same part-of-speech from a vocabulary. The created corpus included various forms of obfuscation like same polarity, add/deletion of words, etc. Then, we studied how this corpus could be useful efficiently in the evaluation of Arabic paraphrase detection using deep contextual word embeddings. AraBERT-LSTM based approach outperformed significantly state-of-the-art methods. It alleviated data sparsity and Arabic word semantics achieving the following F1 scores: 89,75% with OSAC and 86.50% SemEval. For future research, one possible way to further improve our system could be to feed AraBERT embedding to other recurrent neural networks like Bi-directional LSTM (Bi-LSTM), or Gated Recurrent Units (Bi-GRU).

References

Akbik A., Bergmann T., Blythe D., Rasul K., Schweter S., and Vollgraf R. 2019. An easy-to-use framework for state-of-the-art NLP, Conference of the North American Chapter of the Association for Computational Linguistics (demonstrations).

Aliane A. A., and Aliane H. 2020. Evaluating SIAMESE architecture neural models for Arabic textual similarity and plagiarism detection, 4th International Symposium on Informatics and its Applications (ISIA), M'sila, Algeria: 1-8.

Ayata D., Saraclar M., and Ozgur A. 2017. BUSEM at SemEval-2017 Task 4 sentiment analysis with word embedding and long short term memory

RNN approaches, 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, Canada: 777–783.

Babi K., Martincic-Ipsi S., and Meštrovic A. 2020. Survey of neural text representation models, Information, volume.11: 1-32.

Gharavi E., Bijari K., Zahirnia K., and Veisi H. 2016. A deep learning approach to Persian plagiarism detection, FIRE (Working Notes): 1-6.

Haffar N., Hkiri E., and Zrigui M. 2020. Enrichment of Arabic TimeML corpus, International Conference on Computational Collective Intelligence (ICCCI), Da Nang, Vietnam: 655–667.

Haffar N., Ayadi R., Hkiri E., and Zrigui M. 2021. Temporal ordering of events via deep neural networks, 16th International Conference on Document Analysis and Recognition (ICDAR), Lausanne, Switzerland: 762-777.

Hambi M., and Benabbou F. 2019. A new online plagiarism detection system based on deep learning, International Journal of Advanced Computer Science and Applications (IJACSA), 11(9): 470-478.

Hamza A., En-Nahnahi N., and Ouatik S. E. 2020. Contextual word representation and deep neural networks-based method for Arabic question classification, Advances in Science, Technology and Engineering Systems Journal, 5(5): 478-484.

He H., Gimpel K., and Lin J. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks, Conference on Empirical Methods in Natural Language Processing (EMNLP): 1576–1586.

Hindocha E., Yazhiny V., Arunkumar A., and Boobalan P. 2019. Short-text semantic similarity using GloVe word embedding, International Research Journal of Engineering and Technology (IRJET), volume. 6: 553-558.

Iqbal A., Sharif O., Hoque M. M., and Sarker I. H. 2021. Word embedding based textual semantic similarity measure in Bengali, Procedia Computer Science, volume.193: 92–101.

Karaoglan D., Kisla T., and Metin S. K. 2016. Description of Turkish paraphrase corpus structure and generation method, International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Turkey: 208-217.

Liu L., Yang W., Rao J., Tang R., and Lin J. 2019. Incorporating contextual and syntactic structures improves semantic similarity modeling, Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): 1204–1209.

Mahmoud A., and Zrigui M. 2017. Semantic similarity analysis for paraphrase identification in Arabic texts, 31st Pacific Asia Conference on Language, Information and Computation (PACLIC), Philippine: 274-281

Mahmoud A., and Zrigui M. 2019a. Sentence embedding and convolutional neural network for semantic textual similarity detection in Arabic language, Arabian for Engineering and Science Journal, volume. 44: 9263-9274.

Mahmoud A., and Zrigui M. 2019b. Similar meaning analysis for original documents identification in Arabic language, International Conference on Computational Collective Intelligence (ICCCI), Hendaye, France: 193–206.

Mahmoud A., and Zrigui M. 2021a. Arabic semantic textual similarity identification based on convolutional gated recurrent units, International Symposium on INnovations in Intelligent Systems and Applications (INISTA), Kocaeli, Turkey: 1-7.

Mahmoud A., and Zrigui M. 2021b. Hybrid attention-based approach for Arabic paraphrase detection, Applied Artificial Intelligence: 1-16.

Mahmoud A., and Zrigui M. 2021c. Semantic similarity analysis for corpus development and paraphrase detection in Arabic, International Arab Journal of Information Technology (IAJIT), volume. 18: 1-7.

Mahmoud A., and Zrigui M. 2021d. BLSTM-API: Bi-LSTM recurrent neural-based approach for Arabic paraphrase identification, Arabian for Science and Engineering, volume. 46: 4163-4174.

Meddeb O., Maraoui M., and Zrigui M. 2021. Arabic text documents recommendation using joint deep representations learning, Procedia Computer Science, 192(1): 812-821.

Meshram S., and Kumar M. A. 2021. Long short-term memory network for learning sentences similarity using deep contextual embeddings, International Journal of Information Technology, 13(4): 1633–1641.

Mikolov T., Sutskever I., Corrado G., and Dean J. 2014. Distributed representations of words and phrases and their compositionality, arXiv:1405.4053.

Mohammed S. M., Jacksi K., and Zeebaree S. R. M. 2019. A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms, Indonesian Journal of

Electrical Engineering and Computer Science, 22(1): 552-562.

Nagoudi E. B. A., Khorsi H., Cherroun H., and Schwab D. 2018. A two-level plagiarism detection system for Arabic documents, Cybernetics and Information Technologies, In press, 18(1): 1-17.

Othman N., Faiz R., and Smaïli K. 2021. Learning English and Arabic question similarity with Siamese neural networks in community question answering services, Data and Knowledge Engineering (In press): 1-26.

Peinelt N., Nguyen D., and Liakata M. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection, 58th Annual Meeting of the Association for Computational Linguistics: 7047–7055.

Pennington J., Socher R., and Manning C. 2014. GloVE: Global vectors for word representation, Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: 1532-1543.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. 2018. Deep contextualized word representations, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana: 2227-2237.

Ranashinghe T., Orasan O., and Mitkov R. 2019. Enhancing unsupervised sentence similarity methods with deep contextualized word representations, International Conference on Recent Advances in Natural Language Procesing (RANLP), Varna, Bulgaria: 994-1003.

Sghaier M. A., and Zrigui M. 2020. Rule-based machine translation from Tunisian dialect to modern Arabic standard, Procedia Computer Science, volume. 196: 310-319.

Shao Y. 2017. Hcti at semeval-2017 task 1: use convolutional neural network to evaluate semantic textual similarity, 11th International Workshop on Semantic Evaluation (SemEval-2017): 130– 133.

Tai K. S., Socher R., and Manning C. D. 2015. Improved semantic representations from tree-structured long short-term memory networks, arXiv preprint arXiv:1503.00075.

Veisi H., Golchinpour M., Salehi M., and Gharavi E. 2022. Multi-level text document similarity estimation and its application for plagiarism detection, Iran Journal of Computer Science: 1-13.

Yalcin K., Cicekli I., and Ercan G. 2022. An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding, Expert Systems with Applications, volume. 197.

HMIST: Hierarchical Multilingual Isometric Speech Translation using Multi-Task Learning Framework for Automatic Dubbing

Nidhir Bhavsar[$], **Aakash Bhatnagar**[&], **Muskaan Singh**[#] and **Petr Motlicek**[#]
[$]University of Potsdam, Potsdam, Germany
[&] Boston University, Boston, Massachusetts
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
bhavsar@uni-potsdam.de, aakash07@bu.edu,
(msingh,petr.motlicek)@idiap.ch

Abstract

In this paper, we present an approach and impact of isometric neural machine translation on the automatic dubbing process. The length of generated isometric translated sentences ranges within a ±10% of the source text. We propose a hierarchical and multilingual approach toward generating isometric translation via publicly available MUST-C, WMT, and IIT-Bombay(en-hi) datasets. Our experiments use namely, German(de), French(fr), Russian(ru), Italian(it), and Hindi(hi) languages. Additionally, we implement a paraphrasing module with Opusparcus(fr,de,ru), PAWS-X(fr,de) and Topaco(ru) datasets for German, French, and Russian languages to enhance the vocabulary and maintain the isometric constraints. In performance analysis, we report the average length range of source to translation, 55.15% for all languages, while ru exhibits the highest with 62.475% and a relative improvement of 23.04% from the baseline OPUS-MT model.

1 Introduction

Isometric translation is a relatively new concept in neural machine translation. As video content reaches worldwide, it becomes crucial to localize it for different regions. One of the major problems faced while translating media content is the synchrony between the translated output and the visual content. This problem mainly occurs due to a considerable variation in vocabulary between different languages. [1] state that the ideal length of the generated output should be within ±10% range of the source length. The recent machine translation models do not have any parameters to control the length of the output sequences.

To solve the problem mentioned earlier, we fine-tune different pre-train language models using the prompt engineering method. The initial step in all our methods is to identify the appropriate prompt. We use the approach described in [2] to generate prompts while training. Prompt engineering is an efficient way to perform transfer learning while fine-tuning a model.

In this paper, we present, an empirical analysis of our different translation and paraphrasing models. In our approach, the best performing translation model is OPUS MT and the most efficient paraphrasing model is mBART [3]. However, we train the paraphrasing model only for German, Russian and French because of the limited number of languages supported by paraphrasing datasets. We use a combination of Opusparcus and Topaco datasets for Russian, and Opusparcus and PAWS-X datasets for French and German. As per our knowledge, while writing this paper, we cannot find a standard paraphrasing corpus for Italian and Hindi languages.

2 Background

In this section, we further explain neural machine translation in section 2.1, and its corresponding controlling output length in section 2.2 with lexical and length constraint.

2.1 Neural Machine Translation

For a language pair with parallel data as 1, an MT model parameterized with θ, trains to maximize likelihood on the training sample pairs 2.

$$\mathcal{D} = \{(s_i, t_i) : i = 1, \ldots, N\} \tag{1}$$

$$L(\theta) = \theta \arg\max \sum_{i=1}^{N} \log p\left(t_i \mid s_i, \theta\right) \tag{2}$$

2.2 Controlling output length of MT

Several attempts have been made to control the output length attributes, this includes user preference for desired length summarization [4] or using of multiple extractive summarization algorithms for strict length constraints [5], use of side-information [6] or source text involvement and formality [7] [8]. There can be 2 major approaches for constraining sequence length using MT: i) lexically constrained translation, ii) length constrained translation.

2.2.1 Lexically constrained MT

This section includes lexical integration of length-constraint in NMT, either via constrained training or decoding. [9] replaced recognized entities (URL and number) with place-holders which are then detokenized during post-processing. [10] employed a transformer model, augmenting source phrases with target translations to maintain translation consistency while also allowing the machine to learn lexicon translations by duplicating source-side target terms. On the contrary, [11] leverage the effectiveness of Levenshtein Transformers by injecting terminology constraints at inferences time without any significant impact on decoding speed while also mitigating the re-training procedure.

2.2.2 Length constrained MT

[1] injects length control information via the positional encodings of the self-attention, thus enriching the input embedding in source and target with positional information. [12] extend this approach by computing the distance from every position to the end of the sentence, which is further summed with input embedding in the decoder network. [1] combines the methods for biasing the output length by i) conditioning the output to target-source length ratio and ii) enriching the decoder input with relative length embeddings computed according to the desired target string length. [13] involves translating sentences in source language containing pause marker information and integrates verbosity control of phrases between consecutive pause markers.

3 Proposed Methodology

Our methodology incorporates the approach presented by [2]. As shown in figure 1, our model architecture comprises two main components: 1) Translation module and 2) Paraphrasing Module. We experiment with multiple models for both translation and paraphrasing. Our best approaches consist of OPUS MT [14] as the translation model and mBART paraphrasing model. The intuition behind paraphrasing is to increase the vocabulary of the target language. As the primary purpose of isometric translation is to control output length, the paraphrasing model, when used with prompt engineering, helps us adhere to this constraint. For example, the sentence "The name of this person is John" can also be written as "This is John." This example is a precise instance of how paraphrasing can vary the length of generated translated sentences. It is only possible when the model can understand the semantics of a sentence and have the vocabulary to rewrite it. The following two subsections will further elaborate on our two modules.

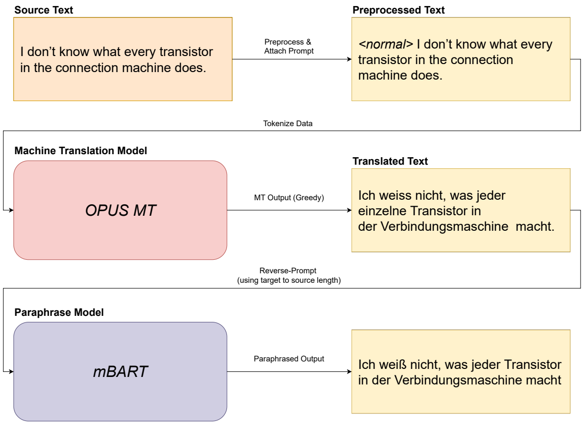

Source Text
I don't know what every transistor in the connection machine does.
Preprocess & Attach Prompt
Preprocessed Text
<normal> I don't know what every transistor in the connection machine does.
Tokenize Data
Machine Translation Model
OPUS MT
MT Output (Greedy)
Translated Text
Ich weiss nicht, was jeder einzelne Transistor in der Verbindungsmaschine macht.
Reverse-Prompt (using target to source length)
Paraphrase Model
mBART
Paraphrased Output
Ich weiß nicht, was jeder Transistor in der Verbindungsmaschine macht

Figure 1: This is the system pipeline of our best-proposed method. After the first step of attaching length prompts, we pass the data from our fine-tuned translation model(OPUS-MT). Further, we pass these translated sentences through a paraphrasing model(mBART) for de,fr,ru.

3.1 Machine Translation

As mentioned earlier, we use OPUS MT for translation. OPUS MT is trained on 108 different languages and performs very efficiently on most languages. Figure 2 describes the general architecture of the OPUS MT model. The OPUS MT model is trained using the Marian NMT framework [15] and constitutes 6-self attentive layers in both encoder and decoder network with 8 attention heads in each layer. However, using baseline OPUS MT does not yield ideal results, as evident in table 4. To comply with the length constraints, we make use of prompt engineering methods and leverage the flexibility of multi-task learning. Our model learns which sentences fall within the normal, short, and long range through prompts. While testing, we add a normal prompt to all input sentences so that model

language	model	dataset	BLEU Score	BERT Score	Length Ratio	Length Range
de	OPUS-MT	MuST-C + WMT	**42.3**	**0.85**	**1.087**	49.81
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	29.1	0.83	1.04	50.55
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	29.9	0.83	1.05	**51.95**
it	OPUS-MT	MuST-C	**34**	**0.84**	**1.045**	**57.032**
fr	OPUS-MT	MuST-C	**44.8**	**0.87**	1.08	49.6
	OPUS-MT+ MT5	MuST-C	42.3	0.85	1.12	51.3
	OPUS-MT + MT5	MuST-C	38	0.86	1.11	46.4
	OPUS-MT + few short mBART	MuST-C + Opusparcus + Paws-X	40.9	0.85	**1.03**	57.33
	OPUS-MT + mBART	MuST-C + Opusparcus + Paws-X	41.2	0.85	1.04	**61.81**
ru	OPUS-MT	MuST-C + WMT	**22.7**	**0.84**	**1.005**	54.517
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	20.8	0.82	0.95	58.934
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	21.7	0.83	0.967	**62.475**
	MT5	MuST-C + WMT	5.6	0.76	0.732	19.3
hi	OPUS-MT	IITB-En-hi	**11.9**	**0.84**	**0.941**	**42.521**

Table 1: Evaluation scores of various experiments. In this table, we state the language-wise experiments along with the datasets used

Source - Target Langauge	Total Instances	Avg. Source Length	Avg. Target Length	Length Ratio	Length Range %
MuST-C					
en-fr	275K	101.78	112.31	1.141	37.65
en-de	229K	100.77	108.84	1.319	36.93
en-it	253K	103.97	108.2	1.076	47.66
en-ru	229K	104.25	102.14	1.044	43..212
IIT-B Corpus					
en-hi	1.65M	74.81	72.92	1.043	46.95
WMT					
en-de	4.5M	138.26	152.45	1.204	28.12
en-ru	2.5M	107.33	98.75	1.18	38.29
Tapaco					
ru-ru	29K	26.38	26.35	1.025	49.45
Opusparcus					
ru-ru	150K	15.81	15.84	1.059	54.948
fr-fr	940K	19.3	19.31	1.079	39.34
de-de	590K	19.63	19.64	1.074	39.33
PAWS-X					
fr-fr	940K	120.94	122.32	1.004	82.61
de-de	50K	119.02	118.24	1.003	85.27

Table 2: Dataset Statistics

generates isometric output.

3.2 Paraphrasing & Length Correction

This module significantly improves our score concerning the isometric constraints. As mentioned above, the main goal of applying the paraphrasing module is to enhance the vocabulary to write sentences with similar meanings in different ways. After exhaustive experimentation, we find that the mBART model is most suitable for paraphrasing. We chose a multilingual version of BART [3] because of its auto-encoding capabilities, which allows it to fully comprehend the language of the text it is parsing, thus making it the best fit for paraphrasing tasks. mBART is a model for text generation in different languages. As seen in table 2, it is evident that the paraphrasing module improves the length ratio and length range significantly.

This module also implies a few-short learning approach via prompt engineering as described by [16].

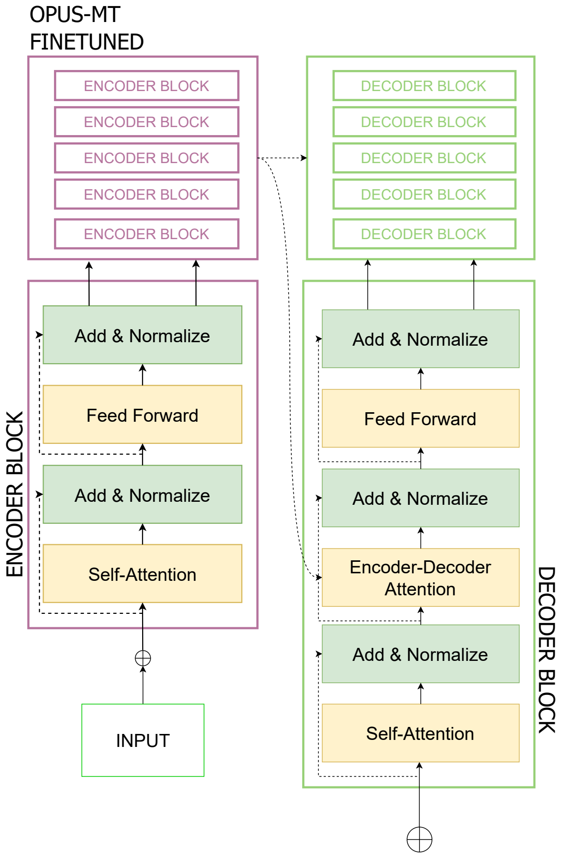


Figure 2: A detailed architecture diagram of our translation model OPUS-MT

The prompts while training are decided similarly to the translation module. However, the difference here is in selecting prompts while making inferences. Given that this model follows the translation module, we slightly employ a different methodology for choosing the prompts for the predictions. We use the paraphrasing module to shorten the long translated outputs or lengthen the short translated outputs. Equation 3 represents the process of finding the appropriate prompt for paraphrasing prediction.

$$f(x) = \begin{cases} \textit{long}, & LR < 0.95 \\ \textit{short}, & LR > 1.10 \end{cases} \tag{3}$$

In equation 3 LR is computed by the length of generated translated text divided by the source text. We do not apply paraphrasing to the translated sentences under the normal range. The paraphrasing module's reverse prompts improve our results and indicate that few short learning is performing as expected. This also indicates that the mBART model successfully understands the length constraint during paraphrasing.

4 Experimental Setup

In this section, we describe dataset details in section 4.1, hyper-parameter settings in Section 4.2 and training procedures in Section 4.3

4.1 Dataset

We implement different datasets for translation and paraphrasing module. As shown in table 1, for machine translation, we use the Multilingual Speech Translation Corpus (MuST-C) [17] for translating majority of our source languages. We also use the Statistical Machine Translation Dataset (WMT) [18] for German (de) and Russian (ru). Additionally for translating Hindi (hi) we use the IIT-B English-Hindi Corpus [19]. Next, we use a combination of Opusparcus [20] and PAWS-X [21] datasets for most of our Paraphrase training tasks, However, due to unavailability of PAWS-X dataset for Russian (ru), we utilize the Tapaco dataset [22] which is a sub-extracted paraphrase corpus derived from the Tatoeba database [23].

4.2 Hyper-parameter Settings

We used 4 Tesla V100-PCIE GPU for all experiments with a memory size of 32510 MiB each. Due to resource constraints, we train each of our models for 1 epoch with a batch size of 32. We apply a learning rate of 2×10^{-5} with a weight decay of 0.01. We implement the AdaFactor optimizer [24], which internally adjusts the learning rate based on the scale parameter and relative/warmup steps.

4.3 Training

In this section, we will discuss details of all experiments performed. In [2], the hierarchical approach was implied only on the French MUST-c dataset with OPUS-MT and MT5 [25] models. This paper extends that approach to five different languages and includes mBART in the paraphrasing module. Our results stand out, and the mBART model performs better than MT5 as a paraphrasing model. This paper also uses reverse-prompt in the paraphrasing module, significantly impacting the results.

We employ prompt engineering on OPUS-MT and MT5 translation models because languages like Italian (it) and Hindi (hi) lack a standard paraphrase dataset. There are very few MT models that support en-hi translation, we utilize only OPUS-MT for this task.

We also implement a singleton MT5 model that performs translation and paraphrasing of 5 supported languages using the prompt engineering method. We utilize the MT5 model for this singleton approach, one of the most optimized multi-task learning models. We use two additional prompts in this approach: 1) Translation and 2) Paraphrasing. The translation prompt signifies that the model will translate the given input, and the paraphrasing prompt signifies that the model will generate isometric sentences from the translated sentences. Further, these two prompts were combined with length prompts. In this model, we use the MUST-C dataset for translation and PAWS-X and Topaco for paraphrasing. However, the sentences generated by this MT5 model are very absurd. After our analysis, we find that the model is mixing up different languages. One possible reason for this is the small dataset size (approx 200K) for each language.

In our experimental setup, we adopt a prompt-based few-shot learning strategy for the paraphrasing task. The model utilizes a small sample of the training dataset(approx 500) and then tries to integrate the derived model with the predictions obtained from the MT model. The comparative scores achieved by assessing using the same technique are listed in table 1. The few-shot model can constrain the output length adequately while also preserving the semantical aspects of the MT. We employ the few-shot learning strategy to train the pre-trained mBART model like the main paraphrase module. Instead of facilitating downstream fine-tuning via pre-training on different corpora, we focus on using

Source Text	Translated Text	Reverse-Pormpt	Paraphrased Sentence	CL Length	BERT Score
I don't know what every transistor in the connection machine does.	Ich weiss nicht, was jeder einzelne Transistor in der Verbindungsmaschine macht.	Short	Ich weiß nicht, was Jeder Transistor in der Verbindungsmaschine macht.	70	0.964
		Normal	Ich weiß nicht, was der einzelne Transistor in der Verbindungsmaschine macht.	77	0.951
		Long	Ich weiß nicht, was der einzelne Transistor in der Verbindungsmaschine macht.	77	0.951
Say, "Please repeat that process." Score them again.	Sag: "Bitte wiederholen Sie diesen Vorgang. " Zählen Sie sie noch einmal.	Short	Bewerte sie nochmal	20	0.749
		Normal	Sag, "Bitte wiederhole diesen Prozess." Bewerte sie.	52	0.979
		Long	Sag, "Bitte wiederhole diesen Prozess" und bewerte sie nochmal.	63	0.908

Table 3: In this table the first & second column represents the source & the generated translated text respectively. The third column shows the value of reverse-prompt that we append on the translated output generated by our model. fourth column represents the paraphrased text generated by our paraphrasing module. CL is the character length of the paraphrased sentences.

accessible data samples to perform few-shot learning. We utilize the prompt-engineering techniques to extend the pre-trained model's performance for the specific task of paraphrasing non-isometric text. We use a similar data configuration for the following source languages: de, fr, and ru.

5 Result and Analysis

For evaluating isometric translation outputs for we use BLEU, BERTScore and length compliance.

- *BLEU* [26] score is a statistical method that evaluates on the basis of n-grams in translated and reference text. Particularly for isometric translation, where the length of translated sentence may vary from the reference text, BLUE score is unable to capture the semantic meaning.

- *BERT score* [27] however, uses pre-trained contextual word embeddings to calculate cosine similarity between translated sentences and reference text. BERT score is the most appropriate option because it can evaluate sentences based on semantics and is comparatively more robust while evaluating short translation sentences.

- *Length Compliance* [28]. is isometric constraint specific evalaution metric which comprises of two measure: (1) Length Ratio and (2) Length Range. Length Ratio is defined as the ratio of source text by generated text. Length Range is defined as the percentage of sentences that falls within ideal span of length ratio 0.90-1.10.

5.1 Evaluation Measures Analysis

Adhering to isometric constraints can negatively affect the derived BLEU score as it depends on the number of characters. The best performing models for the isometric task across each source language have a lower BLEU score, as seen in table 1. This dip in the BLEU score is fairly evident in the languages that use the paraphrase module. The paraphrasing module modulates sentence length to conform to the interchangeable vocabulary. Each of the following language pairs, (en-de, en-fr, en-ru) have a high BLEU score for the baseline OPUS-MT. However, the BLEU score changes abruptly when the paraphrase module is applied, although the value of length compliance metrics improves. Consequently, we recommend the BERT score as a similarity measure since it provides a more precise similarity assessment while taking the semantical

Language	Model	BLEU Score	BERT Score	Length Ratio	Length Range(%)
de	baseline OPUS	**33.1**	**0.84**	1.14	35.74
	finetuned OPUS + mBART	29.9	0.83	**1.05**	**51.95**
it	baseline OPUS	31.3	0.82	**1.037**	54.662
	finetuned OPUS	**34**	**0.84**	1.045	**57.032**
fr	baseline OPUS	**45.4**	**0.86**	1.149	35.41
	finetuned OPUS + mBART	41.2	0.85	**1.04**	**61.81**
ru	baseline OPUS	20.4	0.83	**1.001**	50.776
	finetuned OPUS + mBART	**21.7**	**0.83**	0.967	**62.475**
hi	baseline OPUS	9.9	0.83	0.844	31.911
	finetuned OPUS	**11.9**	**0.84**	**0.941**	**42.521**

Table 4: Comparison of our language-wise best performing systems with the baseline OPUS-MT models

component of translations into account.

5.2 Comparison with Baseline OPUS

We see a significant difference in the length compliance metrics when we compare our results with the pre-trained OPUS-MT model. As depicted in table 4, our best performing models have shown improvement compared to the respective baselines OPUS models. In contrast to the baseline OPUS-MT models, our models can provide length-controlled outputs. In table 4, it is visible that the BLEU score of the pre-trained OPUS-MT model is better than our models in most of the cases; however, there is no significant difference in the BERT score. These statistics further reinforce our point of using the BERT score as an evaluation measure for isometric translation. Particularly for Russian, our predictions Exhibit a high length range of 62.475% and a Length Ratio of 0.967. Moreover, even though the OPUS en-hi corpus is used for pre-training, OPUS-MT consists of only 24M entries compared to 421.5M for de, 550.7M for fr, 241.4M for it, and 160M for ru, we were still able to improve the BLEU score and BERT score.

5.3 Qualitative Analysis

Table 3 provides two instances that demonstrate our reverse-prompt method. As mentioned in earlier sections, reverse-prompt was designed to assist the paraphrasing module in constructing length-controlled phrases. As shown in the second instance of table 3, when the short prompt is applied, our model ignores the content enclosed within the double-inverted commas. At the same time, the long prompt tries to append extra vocabulary to increase the prediction length. As shown in table 3, all three types of outputs (short text, long text, and normal text) exhibit a similar BERT score except for some extreme cases. This consistency in the BERT score represents that our paraphrasing model maintains the semantic meaning while controlling the length of the generated output. When we applied the reverse-prompt approach to Google's MT5 model, the results were not promising compared to mBART. We believe that mBART is more optimized for paraphrasing tasks and prompt-engineering mechanisms.

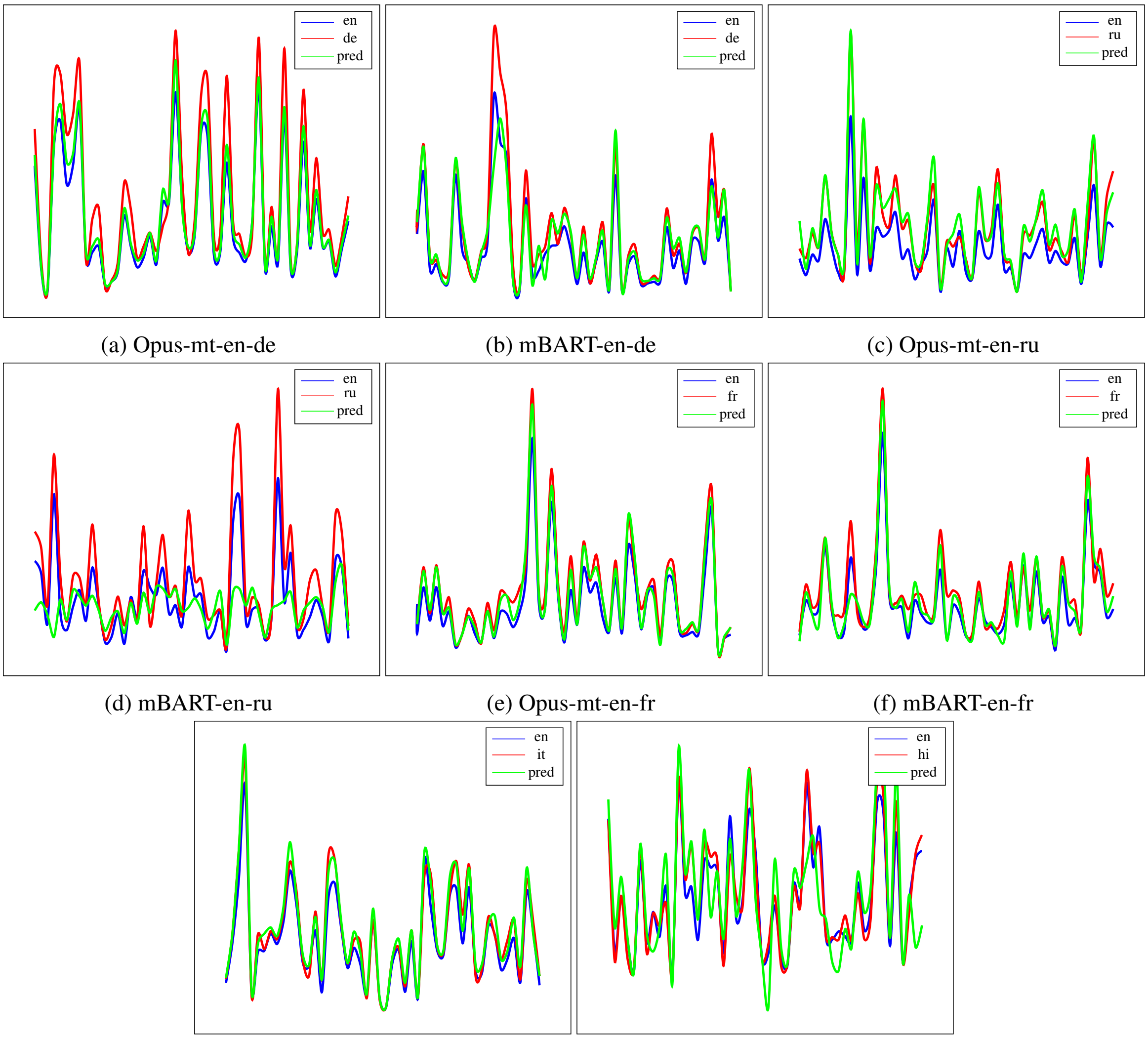
5.4 Automatic Dubbing Analysis

The main purpose of isometric translation is to establish synchrony between the source speech and the translated speech. As stated previously, the ideal range of LR is $0.90 - 1.10$ for source-text translation with isometric constraints so that text-to-speech(TTS) modules can produce a more synchronous output. To analyse this statistic, we use Amazon's Polly(Joanna speaker) [1] and Google's text-to-speech(default speaker) [2] models. Figure 3 shows eight graphs, each representing the time taken by AWS Polly for speaking source language, target language, and our generated isometric translated sentences. We can see that the green and blue lines tend to come together in most cases. This reinforces the claim made by authors of [28] regarding the ideal LR value.

In contrast to AWS Polly, Google's text-to-speech model does not differentiate much between our generated translated outputs and the target outputs. In figure 4 we can see that there is a lot of overlap between the red line and the green line. It is evident that AWS Polly was producing different duration of speech, while Google's Text-to-Speech is not recognizing the difference for the same sen-

[1] https://aws.amazon.com/polly/

[2] https://cloud.google.com/text-to-speech



(a) Opus-mt-en-de

(b) mBART-en-de

(c) Opus-mt-en-ru

(d) mBART-en-ru

(e) Opus-mt-en-fr

(f) mBART-en-fr

(g) Opus-mt-en-it

(h) Opus-mt-en-hi

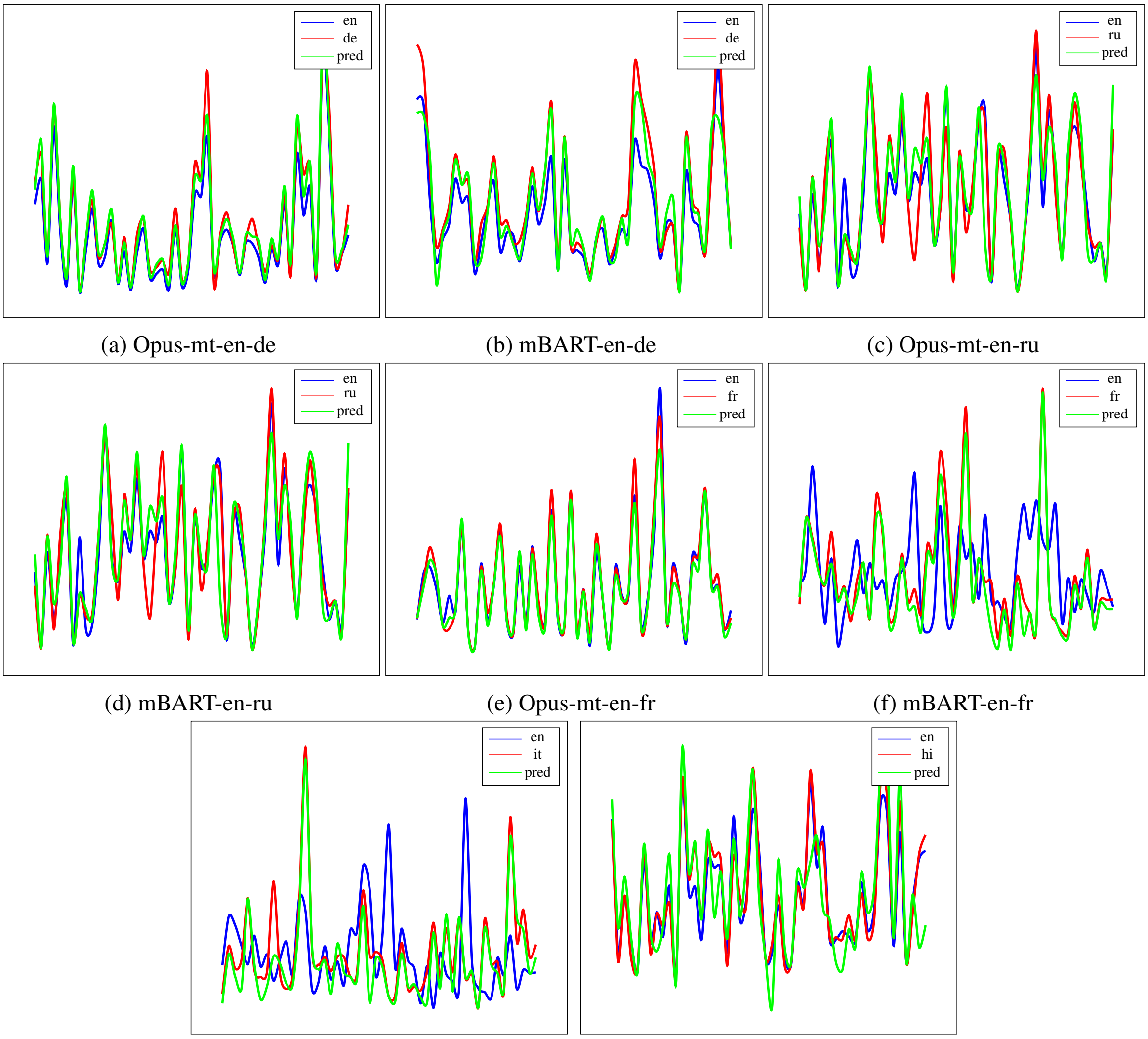
Figure 3: Comparison of time-duration across source, target and prediction for all of the aforementioned models as well as languages (de, fr, ru, it, hi) using Amazon Polly text-to-speech API. Here y-axis represents the duration of speech and *Blue*, *Red*, and *Green* lines show the time duration taken by the model for uttering the source text, target text, and generated isometric translated text.

tences. One possible reason can be that there can be a different ideal range of LR for Google's text-to-speech model to generate isometric outputs. Another reason can be that Google cloud uses wavenet-generated voices[3], which are trained using raw audio samples of actual humans speaking, which lead to a more human-like emphasis and inflection on syllables, phonemes, and words. On the contrary, the AWS Polly produces a more auto-tuned version of voices.

A point worth noting in figure 3 is that although OPUS-MT + mBART model of en-ru achieves the highest length range, the graph of fine-tuned OPUS-MT model seems more convincing and aligned with the source speech. From table 1 we can see that fine-tuned OPUS-MT of ru exhibits the Length Ratio of 1.005, which is extremely close to the ideal value 1. On the other hand, OPUS-MT + mBART achieves the LR of 0.967. After our analysis, we observe that although the LR of OPUS-MT + mBART falls within the isometric constraints, it is shortening most of the sentences.

In our analysis, we also found that the ideal LR can be changed based on the speed of the target language. For example, French is a faster-paced language than English, while German is slower than English. Therefore the value of the ideal LR for generating isometric outputs can be changed. As for a less number of characters, a faster language will match up with English while speaking.

[3] https://cloud.google.com/text-to-speech/docs/basics



(a) Opus-mt-en-de (b) mBART-en-de (c) Opus-mt-en-ru

(d) mBART-en-ru (e) Opus-mt-en-fr (f) mBART-en-fr

(g) mBART-en-fr (h) Opus-mt-en-hi

Figure 4: Comparison of time-duration across source, target and prediction for all of the mentioned models as well as languages (de,fr,ru,it,hi) using Google Cloud's text-to-speech API. Here y-axis represents the duration of speech and *Blue*, *Red*, and *Green* lines show the time duration taken by the model for uttering the source text, target text, and generated isometric translated text.

However, if the number of characters are more in a slower-paced language, it can cope with the English's speed

6 Conclusion & Future Work

In this work, we present a multilingual multitask learning system, which derives relations from the prompt-engineering technique, for fine-tuning the MT models as well as discuss the influence of reverse prompt engineering strategy, which can assist in paraphrasing text by utilizing the reverse prompts obtained using the target to the source character length ratio. We also present a comprehensive study for integrating several neural machine translation models with paraphrase models for source language translations with output length constraints. Additionally, We also investigate the application of a prompt-based few-shot learning technique for paraphrase models extended using the previously trained fine-tuned MT models. However, other enhancements may be incorporated to produce more optimal results. Firstly, there is a shortage of generalized isometric data, limiting the ability to evaluate the MT predictions for statistical metrics such as the BLEU score while also imposing significant constraints on training models for downstream isometric tasks. Next, Our research on the singleton system reveals that existing state-of-the-art multilingual models lack the ability to generalize to the use-case of multilingual MT tasks. Although, a generalization might be added by inferring the MBart model, which can assess language distinction. However, the tokenization criteria, which prevents the simultaneous usage of many target languages, poses a barrier. This could be overcome by utilizing a reasonably distributive word tokenizer. Finally, a lot of improvements can be employed to the current output length control techniques. A combination of positional encoding via self-attention and prompt engineering technique could be employed to signify a more robust isometric MT. Additionally, We can even evaluate the systems predictability for various text-to-speech model thus creating a more diversified length compliance metrics enhanced for Automatic Dubbing.

7 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. Controlling the output length of neural machine translation. *CoRR*, abs/1910.10408, 2019.

[2] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *ACL*, 2022.

[3] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020.

[4] Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. *CoRR*, abs/1711.05217, 2017.

[5] Yashar Mehdad, Amanda Stent, Kapil Thadani, Dragomir Radev, Youssef Billawala, and Karolina Buchner. Extractive summarization under strict length constraints. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3089–3093, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[6] Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. Improved neural machine translation using side information. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 6–16, Dunedin, New Zealand, December 2018.

[7] Xing Niu and Marine Carpuat. Controlling neural machine translation formality with synthetic supervision. *CoRR*, abs/1911.08706, 2019.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540, 2016.

[10] Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. Lexically constrained neural machine translation with levenshtein transformer. *CoRR*, abs/2004.12681, 2020.

[12] Sho Takase and Naoaki Okazaki. Positional encoding to control output sequence length. *CoRR*, abs/1904.07418, 2019.

[13] Derek Tam, Surafel Melaku Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Prosody-aware neural machine translation for dubbing. *CoRR*, abs/2112.08548, 2021.

[14] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[15] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344, 2018.

[16] Chujie Zheng and Minlie Huang. Exploring prompt-based few-shot learning for grounded dialog generation, 2022.

[17] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[19] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT bombay english-hindi parallel corpus. *CoRR*, abs/1710.02855, 2017.

[20] Mathias Creutz. Open subtitles paraphrase corpus for six languages. *CoRR*, abs/1809.06142, 2018.

[21] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828, 2019.

[22] Yves Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association.

[23] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, 2018.

[24] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235, 2018.

[25] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[27] A. A. Vetrov and E. A. Gorn. A new approach to calculating bertscore for automatic assessment of translation quality. *CoRR*, abs/2203.05598, 2022.

[28] Surafel Melaku Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Isometric MT: neural machine translation for automatic dubbing. *CoRR*, abs/2112.08682, 2021.

Bio-Medical Multi-label Scientific Literature Classification using LWAN and Dual-attention module

Deepanshu Khanna[$], **Aakash Bhatnagar**[%], **Nidhir Bhavsar**[&], **Muskaan Singh**[#] and **Petr Motlicek**[#]
[$] ECED, Thapar Institute of Engineering and Technology, India
[%] Boston University, Boston, Massachusetts
[&]University of Potsdam, Potsdam, Germany
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
dkhanna_be19@thapar.edu, aakash07@bu.edu,
bhavsar@uni-potsdam.de, (msingh,petr.motlicek)@idiap.ch

Abstract

An enormous amount of research has been undertaken to overcome the severe impact of COVID-19 pandemic. These scientific findings are being reported in biomedical literature at a significant rate of 10,000 articles/month. In this paper, we tackle automated topic annotation for COVID-19 literature using SPECTER, Bioformer, and PubMedBERT embeddings using Label-Wise Attention Network (LWAN) based Multi-Label Document Classification (MLDC) using Dual-attention module. We also include literature from cardiovascular domain, to generalise our proposed approach. We significantly, achieve 87.71%,72.83% and 79.75% F1-score on LitCovid, Obsumed, WHO-Covid datasets. We release our code-base here https://github.com/Deepanshu-beep/MLDC_LWAN_Attention.

1 Introduction

COVID-19 pandemic, has caused various unexpected challenges to public health and similar line of work. Since its outbreak, there has been a drastic loss in human life, leading to the exponential growth of research and innovation in this field , nearly 20000 articles (till December 2019) have been published.

Due to the information overload, it became a tremendous task for the general public and research professionals to keep up the pace with the latest COVID-19 research. The rate of increase in publications related to the pandemic still continues to increase rapidly today.

Most of the biomedical literature focuses on multiple topics such as treatment, diagnosis, prevention, vaccine etc. These literature are often classified under multiple labels and therefore presents, Multi-Label Document Classification (MLDC) problem at a large scale.

Previously, various researchers have performed thorough experimentation over MLDC. The task of MLDC has been covered in various applications especially in medical domain. One of the most relevant datasets for the medical domain is the MIMIC-III [1], which contains an extensive literature of clinical notes for 16 ICD-9 codes. Similarly, LitCovid is another important dataset that can be used for MLDC, which contains various articles related to COVID-19 corresponding to their 7 unique labels.

Traditional approaches for MLDC included extraction of handcrafted features from documents and then using single or multiple classifiers as in [2]. Earlier works in this domain used Convolutional Neural Network (CNN) [3] and Seq-2-Seq [4]. Authors in [3], later extended their work in [5] by using a 1-dimensional convolutional network to learn text representations and evaluating their approach over 6 datasets. Another work [6] proved the effectiveness of systems involving attention mechanisms by combining Recurrent Neural Networks (RNNs) and self-attention network for MLDC. The paper [6] was one of the few works carrying out a comparison between probabilistic label trees and neural models. Recent top-performing models include methodologies such as Transfer learning, Few and Zero-Shot learning, and Label-Wise Attention Networks (LWANs). Articles [7, 8] experimented with transfer learning using pre-trained language models such as BERT and ELMo, respectively. In [9], authors proposed a Zero-shot attention-based CNN network, which outperformed Zero-shot and Few Shot learning-based methods. Another prominent methodologies included combining LWANs along with BERT [10]. Their work compared LWAN with attention-based RNNs and Hierarchical Attention Network (HAN). Multiple variations of LWANs have been used widely for MLDC, such as CNN-LWAN [11], Z-CNN-LWAN [9]. Hence using LWAN for MLDC became a strong motivation for our work.

In this paper, (1) We tackle the issue of large scale MLDC especially for medical domain by classifying various articles related to COVID-19 and cardiovascular diseases to their classes. (2) We tackle automated topic annotation for COVID-19 and cardiovascular literature using SPECTER [12], Bioformer [13], and PubMedBERT [14] embeddings along with LWAN and Dual-attention module based architecture for Multi-Label Document Classification. (3) We evaluate the performance of the proposed architecture over two COVID-19 articles based databases: LitCovid [15] and WHO-Covid, and another Cardiovascular diseases based database: Ohsumed [16].

2 Methodology

We present the proposed system in Figure 1. Since it is crucial to preserve the contextual importance of the paper to predict its label, we decide to proceed with the classical transformer-based approach, unlike any other neural architecture. Evidently, the transformers-based approach performs eminently compared to CNN or LSTM network-based approaches, which have been discussed later.

Firstly, we create the embeddings for document representations, by concatenating the title and the abstract of the papers using the [CLS] and [SEP] tokens, represented as:

$$< [CLS], title, [SEP], abstract, [CLS] > \quad (1)$$

Secondly, we feed the above input sequence to experiment with three different word embedding methods namely, SPECTER, Bioformer, and PubMedBERT. As these three embedding models are pre-trained over biomedical literature, they were the most suited for this task.

- SPECTER is a pre-trained transformer based pre-trained language model which is an extension of SciBERT [17] and uses citation aware graphs. SPECTER creates high-quality document representations since, unlike other BERT-based models trained on intra-document information, SPECTER was trained on citations as inter-document information. Thus, providing high-quality document-level representations.
- Bioformer is another transformer language model pre-trained over PubMed abstracts and 1 million randomly sampled PubMed central full-text papers. Since LWANs make the architecture computationally expensive, we experimented with Bioformer being a lightweight model with nearly 60% fewer parameters than other BERT-based models. Moreover, it encodes biomedical text efficiently, and the input text length is 20% higher than the other BERT models.
- PubMedBert is also used for experimentation with the proposed network. This state-of-the-art model was pre-trained over PubMed abstracts. The reason behind its exceptional performance is that rather than continuing pre-training of other domain-specific language models over general-domain language models, it was pre-trained from scratch over biomedical data. Hence, behaving exceptionally in various NLP applications.

Finally, we utilise dual attention and LWAN network. A single self attention layer helps model to retain important information in an instance. To further improve this, we implied a dual-attention module that helps in generating relationship between different instances. Dual-attention lead to a significant improvement in results, specially for the classes that have fewer number of instances. In MLDC, LWAN is important because it helps in retaining class-wise information of every instance. Upon retrieving the document representations, we use a Dual-attention module comprising two sequential attention modules. The self-attention module takes word embeddings as input and generates contextualized word embeddings. This attention mechanism is performed by comparing each word with every word in a sentence and recomputing its weights according to the contextual

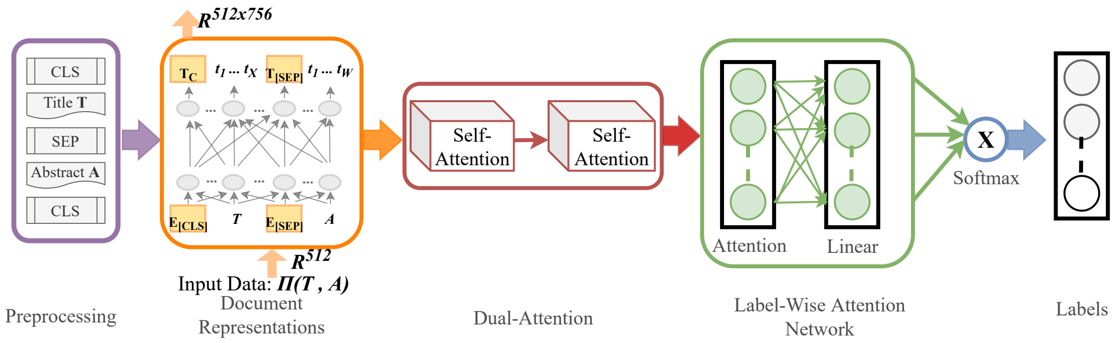


Figure 1: Proposed Methodology

relevance. The attention scores are calculated using three unique vectors: Key (K), Value (V), and Query (Q). The whole self-attention block can be divided in three sub-tasks:

- Calculating dot product similarity between the Query and Key vector *(Q . K_i)*. Upon computing the dot product, we determine the alignment scores that depict the semantic similarity of the pairs. Hence, we get to know the words that require higher attention scores.
- Further, these alignment scores are normalized using a non-linear activation function: Softmax, denoted by a_i.
- Finally, the total attention of words is calculated by multiplying the weights obtained (a_i) and the original document representations embeddings (V_i). The sum of these weighted scores is the output of the self-attention module as shown below:

$$\text{Attention}(Q, K, V) = \sum_i a_i V_i \tag{2}$$

Equation 2: Where: a is Attention score vector, V is the vector corresponding to the embeddings for document representation of Query (Q) and Key (K) vectors.

Using the double self-attention mechanism focuses more on the input information in a sentence and generates better attention vectors as compared to a single self-attention module. Since the self-attention mechanism generates embeddings with relationship amongst the input instances only, which is even more precise when using the Double-attention mechanism, it disregards the output completely. Therefore, to overcome this limitation, we introduce the use of Label-Wise-Attention-Network (LWAN) in our model that provides the attention for each label to classify in the dataset.

LWANs play a significant role in tasks related to multi-label prediction. They assign particular attention scores for every attention node corresponding to the output label while focusing on words with higher attention in input as well. Here, the attention scores are calculated by applying the same attention mechanism for the unique labels. The mathematical procedure for calculating attention scores for LWAN is also show below:

$$z_{i,l} = w_{a,l} h_i + b_{a,l} \tag{3}$$

$$\alpha_{i,l} = \frac{e^{z_{i,l}}}{\sum_{j=1}^{N} e^{z_{j,l}}} \tag{4}$$

$$s_l = \sum_{j=1}^{N} \alpha_{j,l} h_j \tag{5}$$

$$\beta_l = w_{f,l} s_l + b_{f,l} \tag{6}$$

$$p_l = \frac{e^{\beta_l}}{\sum_{r=1}^{L} e^{\beta_r}} \tag{7}$$

Equation 3-7: Where: $w_{a,l}$, $b_{a,l}$ correspond to weight and bias vector and h_i is the hidden LSTM representation of the i^{th} word. $\alpha_{i,l}$ are the attention weights obtained after normalizing $z_{i,l}$ using Softmax function. Further we obtain the whole sentence representation by computing the weighted average of h_i. Finally, we obtain the confidence score of l_{th} label

	LitCovid	**Ohsumed**	**WHO-Covid**
No. of unique abstracts	31199	34389	169292
No. of categories	7	23	19
Average sentences	10	9	10
Average tokens for input	212	182	216
Total no. of tokens	6618763	6248274	36605199

Table 1: Statistics of LitCovid, Ohsumed and WHO-Covid databases.

Model	F1-score
Team BJUT-BJFU @ Biocreative	78.47
KimCNN	83.45
LSTM	83.95
LSTM$_{reg}$	84.05
XML-CNN	84.2
SPECTER-Dual Attention-LWAN	87.13
Bioformer-Dual Attention-LWAN	87.64
PubMedBERT-Dual Attention-LWAN	87.71
Bioformer	**88.75**

Table 2: F1-score performance over LitCovid database for different embeddings used for document representations and comparison with various models.

Model	F1-score
Sentence2Vec	37.34
LSTM	62.7
HAN	67.0
TextING	69.5
HAN+TextING	70.3
Bioformer-Dual Attention-LWAN	71.17
HGMETA	72.0
PubMedBERT-Dual Attention-LWAN	72.48
SPECTER-Dual Attention-LWAN	**72.83**

Table 3: F1-score performance over Ohsumed database for different embeddings used for document representations.

Model	F1-score
Bioformer-Dual Attention-LWAN	75.61
PubMedBERT-Dual Attention-LWAN	78.87
SPECTER-Dual Attention-LWAN	79.75

Table 4: F1-score performance over WHO-Covid database for different embeddings used for document representations.

denoted by β_l, which is further normalized using Softmax function and denoted as: p_l.

3 Experimental Details

In this section, we present, our experimental settings, with dataset in section 3.1, hyperparameter in section 3.2 and training in section 3.3.

3.1 Datasets

We use three datasets: LitCovid, Ohsumed, and WHO-Covid, to experiment with our proposed architecture and prove its efficiency. A detailed description of the datasets has been given below. Various statistics of the datasets have also been depicted in the Table 1.

- LitCovid is a large corpus of articles published in PubMed about the COVID-19 and SARS-COV-2. The dataset contains in total 31199 unique articles corresponding to 7 classes, namely: Case Report, Diagnosis, Epidemic Forecasting, Mechanism, Prevention, Transmission and Treatment. We used 24960, and 6239 articles for our train and dev set, respectively.

- Ohsumed dataset is a subset of the MEDLINE database containing medical abstracts MeSH categories from 1991. The dataset is aimed to classify 23 cardiovascular disease categories. Ohsumed contains 34389 unique abstracts, out of which we used 27511 and 6878 for the training and validation.

- WHO-Covid Since the global pandemic of COVID-19, WHO has been collecting global literature on COVID-19. This database gets updated regularly from Monday to Friday. The database consists of multilingual content from searches of bibliographic databases, manually searched articles, and various experts referred scientific articles. During the time we scraped the articles, it had in total 169292 articles for the English language that were selected for experimentation.

3.2 Hyperparameter Setup

In our experiments, we analyze the performance of various embeddings for creating document representations. Table 2,3 and 4 show the performance of the proposed model with different embeddings experimented over the following databases.

As discussed earlier, one of the common challenges in MLDC is the imbalanced distribution of labels for training. Initially, we tackled the issue by adopting the weighted Binary-Cross entropy (BCE) loss function, which improved the performance by a slight margin. Later, we assign weight to each class to focus on minority labels as well rather than only on dominant labels. We calculate the weight for each class using the below formula:

$$W_i = \frac{C_m}{C_i} \tag{8}$$

Equation 8: Where: C_i represents the count of the i^{th} label, C_m represents the count of the most dominant label and W_i corresponds to the weight of the label.

3.3 Training

For the LitCovid database, we trained our model over the original train and dev set released, but for Ohsumed and WHO-Covid databases, we split the data in 80:20 for training and evaluation, respectively. We used the Weighted BCE loss function and trained for ten epochs for LitCovid and Ohsumed databases while for 25 epochs for the WHO-Covid database, due to its large size. We use a learning rate scheduler that changes the learning rate to 2 x 10^{-5} and linearly went down until 0. Other details of the hyperparameters have been shown in Table 5.

3.4 Result and Analysis

In our performance analysis, SPECTER performed best out of the 3 document representations selected. We achieved 87.71, 72.48, and 79.75 macro F1-score for LitCovid, Ohsumed, and WHO-Covid databases. Though the highest score we achieved for the LitCovid database is 87.71 using PubMedBERT embeddings, which is just 0.46% higher than that of SPECTER, hence proving its overall efficiency. Also, we observe that assigning class weights significantly improves the macro F1 scores for rare classes. For instance, we achieved an F1 score of 77.96 for the Epidemic Forecasting label in the LitCovid database, which is the rarest class in it.

The proposed architecture, despite its simple network, performs remarkably. For the LitCovid database, the current top-performing architecture [13] used the Bioformer language model fine-tuned over the database and achieved an F1 score of 88.75, 1.18% higher than the proposed model. We outperform various deep CNN based models: KimCNN [18], XML-CNN [3] along with regularized and unregularized LSTM [19]. Some of the examples predicted by the proposed model have been shown in Table 6. In [20], experimented with four different models: FastText, TextRCNN, TextCNN and Transformer. Their top-performing model achieved an F1-score of 78.47.

For the Ohsumed database, all of our experimented document representations outperform existing models. [21] presented HGMETA that initially extracts fusion embeddings of hierarchical semantics dependence and graph structure in a structured text. Further, they use a hierarchical LDA module and a structured text embedding module to merge the extracted hierarchical features with structured text information. HGMETA being the best performing model on Ohsumed yet, achieved an F1-score of 0.72. Another graph-based method, TextING [22] that treated each document as an individual graph while training, achieved an F1-score of 0.695. On the other hand, HAN [23] that is most commonly used method for structured text classification, uses a RNN-based network along with hierarchical attention mechanism achieved an F1-score of 0.67. Performances of various other baseline methods over Ohsumed database have been shown in Table 3.

4 Conclusion and Future Work

We present our proposed approach using Dual-Attention LWAN method and experiment with three different embeddings namely: SPECTER, Bioformer and PubMedBERT, for document representations. We evaluated the performance of the proposed architecture over three databases: two related to COVID-19 articles and the other one general biomedical articles based. We also observe that despite LWANs focusing just on input data, they perform remarkably when combined with a dual attention module. The only limitation faced during the implementation is the computational expense of LWAN when training upon databases with a large number of labels. As future work, we plan to mitigate this limitation by exploring other novel architectures and methodolo-

Model	Hyperparameters	Number of parameters
SPECTER-Dual Attention-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 4 , 32 epochs: 10 , 25 model: SPECTER warmup proportion: 0.2 dropout probability: 0.1	model: $109M$ Attention: $1.8M$
PubMedBERT-Dual Attention-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 4 , 32 epochs: 10 , 25 model: PubMedBERT (Abstract + Full-text) warmup proportion: 0.2 dropout probability: 0.1	model: $109M$ Attention: $1.8M$
Bioformer-Dual Attention-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 4 , 32 epochs: 10 , 25 model: Bioformer-Cased warmup proportion: 0.2 dropout probability: 0.1	model: $42.5M$ Attention: $786K$

Table 5: Hyperparameter details for he experimental setup, for spectar-dual attention-LWAN, PuBMedBERT-Dual Attention-LWAN and Bioformer-Dual Attention

Article	**Actual**	**Predicted**
Cardiac dysfunction in Multisystem Inflammatory Syndrome in Children	Clinical Practice Guide Observational Study Prognostic Study Risk Factors	Observational Study Risk Factors
I Told You the Invisible Can Kill You": Engaging Anthropology as a Response in the COVID-19 Outbreak in Italy".	Etiology Study Observational Study	Etiology Study Observational Study Risk Factors

Table 6: Examples of predictions by our model (Green color indicates the correct predictions, red color indicates incorrect predictions.

gies along with handling the data imbalance more effectively to create more effective MLDC models.

5 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[2] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning*

and knowledge discovery in databases, pages 437–452. Springer, 2014.

[3] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124, 2017.

[4] Venkatesh Umaashankar and Girish Shanmugam S. Multi-label multi-class hierarchical classification using convolutional seq2seq. In *KONVENS*, 2019.

[5] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, 2021.

[6] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Anthony Rios and Ramakanth Kavuluru. EMR coding with semi-parametric multi-head matching networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2081–2091, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. *arXiv preprint arXiv:2010.01653*, 2020.

[11] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics, 2020.

[13] Li Fang and Kai Wang. Team bioformer at biocreative vii litcovid track: Multic-label topic classification for covid-19 literature with a compact bert model. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, 2021.

[14] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[15] Qingyu Chen, Alexis Allot, and Zhiyong Lu. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540, 2021.

[16] William Hersh, Chris Buckley, TJ Leone, and David Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer, 1994.

[17] Iz Beltagy, Arman Cohan, and Kyle Lo. Scib-

ert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019.

[18] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[19] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Rethinking complex neural network architectures for document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4046–4051, 2019.

[20] Shuo Xu, Yuefu Zhang, and Xin An. Team bjut-bjfu at biocreative vii litcovid track: A deep learning based method for multi-label topic classification in covid-19 literature.

[21] Shaokang Wang, Li Pan, and Yu Wu. Meta-information fusion of hierarchical semantics dependency and graph structure for structured text classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2022.

[22] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, Online, July 2020. Association for Computational Linguistics.

[23] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

An Empirical Comparison of Semantic Similarity Methods for Analyzing down-streaming Automatic Minuting task

Aditya Upadhyay[&], Aakash Bhatnagar[1], Nidhir Bhavsar[$], Muskaan Singh[#] and **Petr Motlicek[#]**
[&] CSED, Thapar Institute of Engineering and Technology, India
[1] Boston University, Boston, Massachusetts
[$]University of Potsdam, Potsdam, Germany
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
adityaupadhyay1912@gmail.com, aakash07@bu.edu,
bhavsar@uni-potsdam.de, (msingh,petr.motlicek)@idiap.ch

Abstract

Automatic Minuting consists of automatically creating minutes from multiparty meeting transcripts. In this paper, we solve two relevant problems of this domain (1) given a pair of meeting transcript and minute, the task is to identify whether the minutes belongs to the transcript. (2) given a pair of minutes, the task is to identify whether the two minutes belong to the same or different meetings. These challenging problems are important as we want to uncover how minutes created by two different persons for the same meeting may differ in content and coverage. The proposed system leverage off-the-shelf semantic similarity techniques which provides with a score of similarity indicating a measure of how close the two text are to each other in meaning. In performance analysis, we broadly formulate three categories with the best performers in each (1) in lexical summarization DOC2VEC (2) in machine learning (3) in deep transformer architectures. We evaluate each of our proposed approaches on the basis of Accuracy. For lexical summarization, Doc2Vec achieves 90% and 51% accuracy, in machine learning, random forest achieves 91% and 85% accuracy and in deep learning, STSB-BERT-Large achieves 94% and 81% accuracy in transcript-minutes and minutes-minutes respectively.

1 Introduction

Due to differences in the style of minuting there are two important challenges,(i) identify if the minutes and the transcript are from the same meeting. (ii) identify if two minutes are from the same meeting (which are taken by different note takers). In this paper, we focus to solve this novel problem and see to what extent this decision can be carried out automatically. The novelty of our research is to *examine the subjectivity associated with the minuting exercise*. Minuting is a challenging task [1], and even more difficult is identify meetings similarity on similar topics with (1) similarity of discussed content and anchor points like named entities e.g. in recurring meetings of the same project on the one hand, and the differences in the style of minuting on the other hand. (2) some minutes do not capture the central points in the meeting because the external scribes did not understand the context correctly and created minutes which miss significant issues discussed in the meeting or are simply too short.

2 Background

Semantic Similarity is a Natural Language Processing (NLP) task that consists of measuring the similarity between two texts in a quantitative manner. Measuring semantic similarity between two texts serves an important role in various NLP tasks such as information retrieval, text summarization, text classification, essay evaluation, machine translation, and question answering [2]. One of the major challenges in the semantic similarity problem is the complex nature of semantics, for instance the words *car* and *gasoline* are definitely more closely related than the words *car* and *bike*, but the latter pair is more similar than the former. Some of the earliest methods used to solve the semantic similarity problem involved embedding words using systems of taxonomy that paired similar words together in trees, such as the WordNet taxonomy. The major issue in this approach was that it relied on the assumption that links in the taxonomy represented similar distances which was not always the case.

3 Related Work

There are various different methods to measure semantic similarity. These include deep learning approaches, machine learning approaches and standard algorithmic approaches. [2] provides a survey of various different methods to semantic similarity along with datasets and semantic distance measures. Within machine learning approaches regression is a popular technique to measure semantic similarity regression is a predictive modelling technique that is used to obtain the relation between the target and the input features. [3] compares various different regression models on the SemEval dataset. They found Boosting to have the best performance when compared to methods such as Bagging, Multi Linear, RPart, Random Forest, and SVM algorithms. [4] proposed using pairwise word interactions in order to find the context based correlation as neural based approaches seem to have difficulty in finding word level similarity. The model performance was texted using news articles, headlines and other such datasets and the model was found to perform very well compared to standard neural networks.

[5] proposed an improved version of the Bi-LSTM model based on the Siamese network architecture. The model was trained on a QA dataset consisting of 100,000 question pairs, 10 fold cross validation was used as a loss metric and the model was found to perform significantly better than other models achieving an accuracy of 84.87%. [6] implemented SVM with CNN and Siamese Recurrent architecture for RNN on a QA dataset. They found the CNN with SVM doesn't correctly assess if the statement has some image and RNN has a Vanishing Gradient problem. [7] also proposed 7 different variants of the RNN architecture using the SICK dataset and the STS2017 dataset. The best performance was achieved by a model that contains a single GRU cell.

Embeddings based neural networks are also widely used to the task of semantic similarity.[8] compares various embeddings based models trained on 1.7 million articles from the PubMed Open Access dataset. These models were tested on a biomedical benchmark (BIOSSES) set that contains 100-sentence pairs. They found Paragraph Vector Distributed Memory algorithm to outperform all other models achieving a correlation of 0.819. [9] compared a CNN model with six other models and used the LIME algorithm to identify the keywords and improve model performance. The other approaches based on deep learning methods [10, 11, 12, 13, 14, 15, 16, 17, 18, 18, 19, 20, 21, 22, 23, 24, 25, 26] proposed by various researchers improve semantic similarity for different applications.

4 Methods

We perform an empirical comparison using off-the-shelf methods of semantic similarity to downstream novel task of determining whether minutes belong to a meeting and whether two sets of minutes belong to a meeting. We formulate two broad categories, namely (1) lexical similarity techniques (2) machine learning based similarity techniques (3) deep transformer architectures.

The lexical similarity methods compares word lengths and character-wise similarity by embedding the contents of the input texts into vectors and then determining the semantic distance between those vectors. In our work we use,

- Bag-Of-Words (BoW) [27] was one of the first methods to embed data for text classification ever developed. It converts a document into a set of words keeping the frequency of each word as a feature in the set. This frequency is used as the embedding for the term.

- Doc2Vec is an implementation of paragraph embedding that was initially proposed in [28]. paragraph embedding uses a log-probability function to obtain the probability of each word in the input text and then uses a function such as softmax to classify the word into a vector. Doc2Vec uses hierarchial softmax to embed the input text.

- Named Entity Similarity [29] is a method that extracts the named entities in a document (real world objects) and embeds them based on the type of entity they correspond to, for instances Apple is an organization while U.K. is a Geopolitical Entity.

- Keyword Similarity Words that seem important or representative of the text are extracted as keywords using the BM25 Ranking Function proposed in [30] to extract the importance of a word. The sets of keywords extracted are then compared to determine similarity.

- Cosine Similarity measures the similarity between two vectors via inner product. It is measured by the cosine of the angle between two vectors and determines whether two text article/documents are pointing in roughly the same direction. For computing the similarity between the text documents we considered using the cosine similarity pairwise metric by sklearn.

$$\text{similarity } = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \quad (1)$$

- Summarized Keyword Similarity, entire texts are summarized using the BM25 Ranking Function to extract important sentences. Keywords are then extracted using the same method to get summarized keywords for each. These summarized keywords are compared to determine similarity.

- Summarization based Named Entity Similarity, entire texts are summarized using the BM25 Ranking Function to extract important sentences. Named entities are then extracted and embeded based on the entity they correspond to. These embeddings are compared to determine similarity.

- SequenceMatcher is a class that is available under the difflib Python package[1]. Main objective behind *SequenceMatcher* is to find the longest contiguous matching sub-sequence LCS with no "irrelevant" elements. Irrelevant are the characters that we don't want the algorithm to match, like blank lines in ordinary text files, etc. This metric does not yield minimal edit sequences, but does tend to yield matches that logically seems appropriate.

- Jaccard Similarity, measure the similarity of two meeting minutes in terms of their context, i.e. how many common words there are compared to the total number of words. Here J is the Jaccard distance calculated via the distinct word present in set A and B.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

These methods pose challenge, to capture the syntactic and semantic meaning of the words in a text and therefore cannot effectively measure semantic similarity. For instance, "I have a weak disposition" and "I often get sick" have the same meaning but due to a lack of similar words would not measure a high semantic similarity using these methods. Additionally, synonyms and ambiguous words (have multiple meanings in different contexts).

In machine learning, we implement Support Vector Machine (SVM) and the Random Forest Classifier.

please add about SVM and random forest

Deep transformer architectures, embed text into vectors to measure semantic similarity. These models are trained on massive corpora of words and take syntactic meaning as well as context into account when embedding sentences. This allows them to make relatively accurate measures of semantic similarity when compared to traditional approaches.

In our work, we use,

- Universal Sentence Encoder is a sentence encoder that was developed by researchers at google in [31]. The encoder was trained on unsupervised data collected from various sources including Wikipedia, various web news and discussion forums. The unsupervised data was augmented with training on supervised data from the Stanford Natural Language Inference (SNLI) corpus [32].

- BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional transformer model proposed in [33]. BERT was trained using a combination of masked language modeling objective and next sentence prediction (NSP) on a large corpus comprising the Toronto Book Corpus and Wikipedia. In our work, we experiment with BERT namely stsb-bert-large [2], stsb-bert-base [3], nli-bert-large [4], nli-bert-large-max-pooling [5], nli-bert-large-cls-pooling [6], nli-

[1] https://docs.python.org/3/library/difflib.html

[2] https://huggingface.co/sentence-transformers/stsb-bert-large

[3] https://huggingface.co/sentence-transformers/stsb-bert-base

[4] https://huggingface.co/sentence-transformers/nli-bert-large

[5] https://huggingface.co/sentence-transformers/nli-bert-large-max-pooling

[6] https://huggingface.co/sentence-transformers/nli-bert-large-cls-pooling

bert-base-max-pooling [7],nli-bert-base [8], nli-bert-base-cls-pooling [9]

- RoBERTa (Robustly optimized BERT approach) [34] uses the same architecture as BERT but modifies the pre-training step. Specifically, RoBERTa is trained with dynamic masking, FULL-SENTENCES without NSP loss, large mini-batches and a larger byte-level BPE. We experiment with stsb-roberta-large[10], stsb-roberta-base[11], nli-roberta-large[12], nli-roberta-base[13].

- DistilBERT (Distilled BERT) [35] is a fast and light variant of BERT. It is trained 40% less parameters than BERT, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark. We use stsb-distilbert-base[14], paraphrase-distilroberta-base-v1[15], nli-distilbert-base[16], nli-distilbert-base-max-pooling[17].

- XLM (cross-lingual Language Model) [36] is a transformer based model that is trained on Next Token Prediction (causal language modeling (CLM) objective), masked language modeling (MLM) objective and a Translation Language Modeling (TLM) object.In our work, we use paraphrase-xlm-r-multilingual-v1 model[18].

5 Experimental Setup

In this section, we describe our experimental setup for the empirical comparison of off-the-shelf methods for our novel down-streaming task. We describe dataset in Section 5.1 and hyperparameter in Section 5.2.

5.1 Dataset Details

The dataset for determining whether minutes belong to a meeting, consists of pairs of transcripts and minutes that are labelled either True or False depending on whether they were derived from the same conversation or not, i.e. True implies the minutes match the transcript and vice versa. The other problem, consists of pairs of minutes that are labelled True of False depending on whether they belong to the same meeting or not. We have used the English and Czech datasets for both tasks. Both these datasets are strongly imbalanced with only around 15% of the pairs belonging to the True class in each case.

5.2 Hyperparameters

We use similar hyperparameters of all the transformer models, with sequence length of 128, word embedding dimension as 1024, drop_out rate of 0.1, hidden_size of 1024, initializer_range as 0.02, intermediate_size of 4096, layer_norm 1e-05 epissilon value, and max_position_embeddings of 514. There are a few more parameters such as pooling type that are different for different models. Some use max pooling while others use mean or CLS pooling. These hyperparameters were picked to tune the models to improve performance. The models were trained on their respective datasets using these hyperparameters

To perform this classification, the similarity values are produced on the embedding produced by a pre-trained model, and then a threshold is used to achieve the binary classification. The pretrained model used is ”bert-base-nli-mean-tokens” provided by hugging face. In this model, BERT-base has been used, which creates the dense vectors containing 768 values. These 768 values contain our numerical representation of a single token — which we can use as contextual word embedding. Some other hyperparameters for this model

[7] https://huggingface.co/sentence-transformers/nli-bert-base-max-pooling

[8] https://huggingface.co/sentence-transformers/nli-bert-base

[9] https://huggingface.co/sentence-transformers/nli-bert-base-cls-pooling

[10] https://huggingface.co/sentence-transformers/stsb-roberta-large

[11] https://huggingface.co/sentence-transformers/stsb-roberta-base

[12] https://huggingface.co/sentence-transformers/nli-roberta-large

[13] https://huggingface.co/sentence-transformers/nli-roberta-base

[14] https://huggingface.co/sentence-transformers/stsb-distilbert-base

[15] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v1

[16] https://huggingface.co/sentence-transformers/nli-distilbert-base

[17] https://huggingface.co/sentence-transformers/nli-distilbert-base-max-pooling

[18] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

include the non-linear activation function (function or string) in the encoder and pooler as 'gelu', the number of attention heads for each attention layer in the transformer encoder as 12 and the standard deviation of the truncated-normal-initializer for initializing all weight matrices as 0.02. After this model produces the embeddings, cosine similarity (measured similarity between two feature vectors by capturing the document's orientation and not the magnitude, unlike the Euclidean distance) is measured, and the final similarity values are produced. Many threshold values are checked to minimize the mismatching of actual binary classifications and the generated binary classifications. The final threshold value is chosen by 0.65.

A pretrained model was used to obtain the sentence embedding. Then a similarity metric was used to get the similarity values, and finally, a threshold value was obtained to get the final binary classification. The pretrained model used is "paraphrase-distilroberta-base-v1', which is a 'DistilBERT-base-uncased'model fine-tuned on a large dataset of paraphrase sentences. This RoBERTa-based sentence representation model has been trained to produce meaningful sentence embedding for similarity assessment and retrieval tasks. It uses a vector length of 768 for the sentence embeddings. Some other hyperparameters for this model include the non-linear activation function (function or string) in the encoder and pooler as 'gelu', the number of attention heads for each attention layer in the transformer encoder as 12 and the standard deviation of the truncated-normal-initializer for initializing all weight matrices as 0.02. After this model produces the embeddings, cosine similarity is measured, and the final similarity values are produced. Many threshold values are checked to minimize the mismatching of actual binary classifications and the generated binary classifications. The final threshold value is chosen by 0.65. The final scores yield an accuracy of 79.8%.

6 Results and Analysis

All the above listed models were tested on the Automin Dataset Minutes-Transcript and Minutes-Minutes using cosine distance as a measure of semantic distance. The testing was carried out on an Nvidia K80 with 2496 CUDA cores operating at 4.1 TFLOPS with 12 GB of primary memory and a hyper-threaded Intel Xeon processor with 2 cores operating at 2.3 GHz. The compute times of each approach can be found in Figure 1.

In our results, we perform quantative evaluation using accuracy and we also vouch for qualitative analysis. The results of the tests on lexical analysis methods can be found in Table 1. It can be observed that Keyword similarity had by far the best performance with Summarization Keyword Similarity being a close second. Summarization based Named Entity Similarity had the best performance on minutes-transcript but performed poorly on minutes-minutes, this disparity can be attributed to the nature of the datasets. The results of the tests performed on the machine learning algorithms can be found in Table 3. The results of the tests on Transformer Based Deep Learning models can be found in Figure 2. In Figure 1 we can observe the computational times for the different deep learning models. Snippets of the datasets for true positive, true negative, false negative and false positive results from the stsb-bert-base model can be found in Figure 3, Figure 4, Figure 5 and Figure 6 respectively.

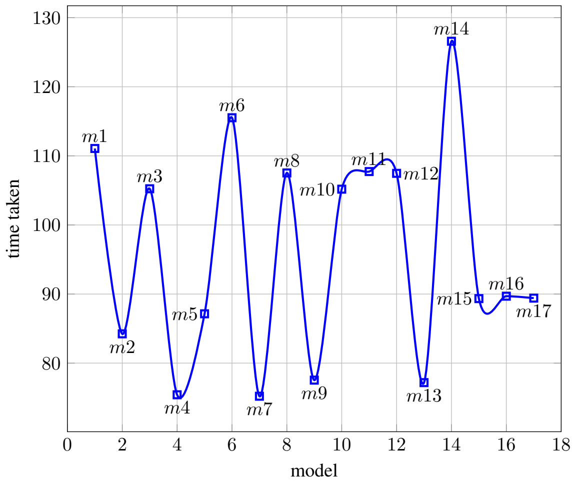


Figure 1: m1: stsb-roberta-large, m2: stsb-roberta-base, m3: stsb-bert-large, m4: stsb-distilbert-base, m5: stsb-bert-base, m6: paraphrase-xlm-rmultilingual-v1, m7: paraphrasedistilrobertabase-v1, m8: nli-bert-large, m9: nli-distilbert-base, m10: nli-roberta-base, m11: nli-bert-large-maxpooling, m12: nli-bert-large-clspooling, m13: nli-distilbert-basemax-pooling, m14: nli-roberta-base, m15: nli-bert-base, m16: nli-bert-base-cls-pooling.

7 Conclusion

Based on the observations drawn from the tests performed, we can conclude that Transformer

Table 1: This table shows the results of all the lexical similarity methods for semantic similarity

Approach	Accuracy (Task B)	Accuracy (Task C)
DOC2VEC	0.9089577851	0.51258137
Named Entity Similarity	0.04519868388	0.05636713945
Keyword Similarity	0.7957055804	0.68533914
Summarization Keyword Similarity	0.523513454	0.4626331746
Summarization based Named Entity Similarity	0.02417613166	0.7510776139
Feature Engineering (Bag of Words)	0.21272047	0.01208951125
Universal Sentence Encoder	0.2644637739	0.4936408219

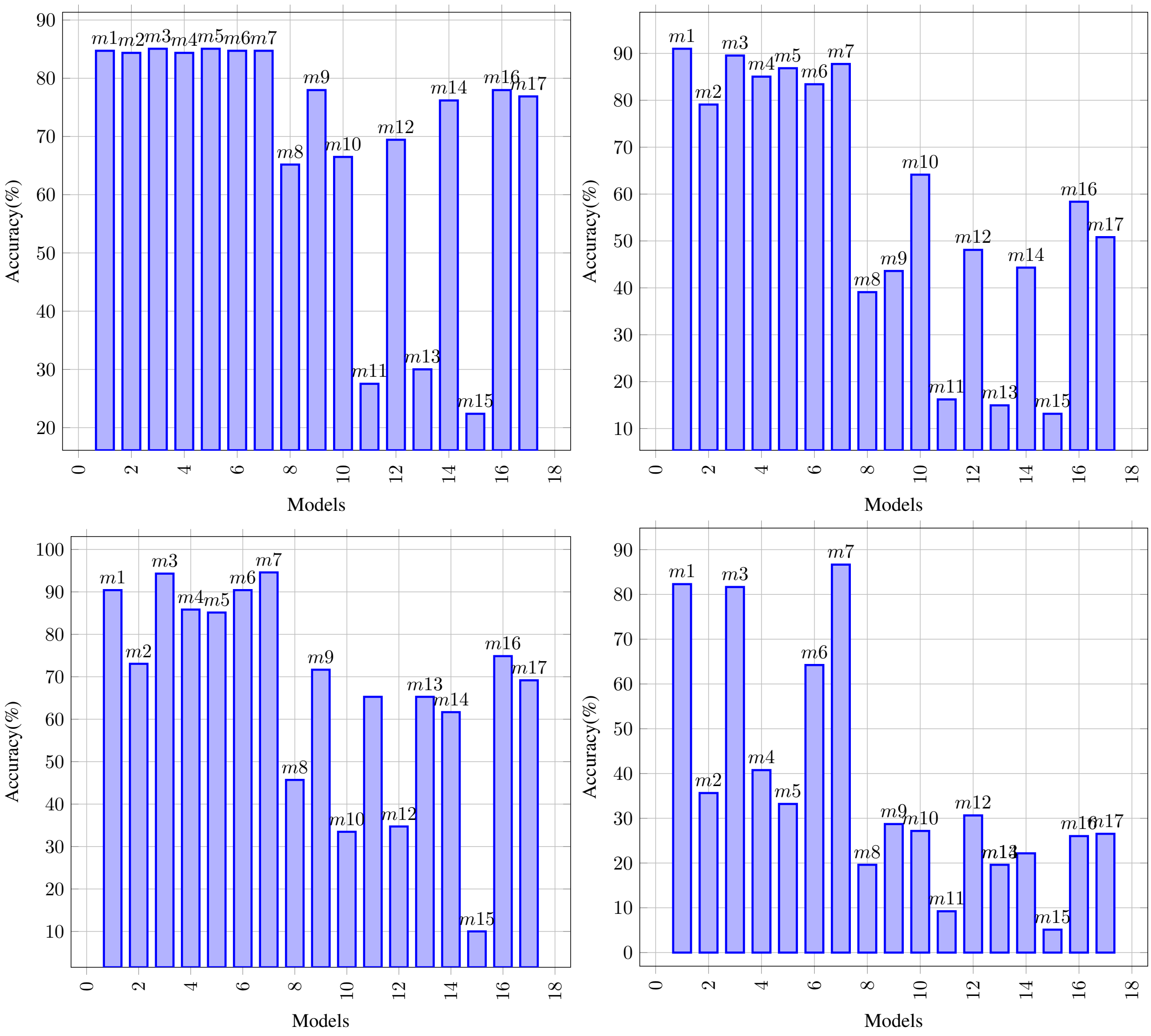


Figure 2: The first row contains the evaluation score of english language, while the bottom two graphs contain evaluation scores of czech m1: stsb-roberta-large, m2: stsb-roberta-base, m3: stsb-bert-large, m4: stsb-distilbert-base, m5: stsb-bert-base, m6: paraphrase-xlm-rmultilingual-v1, m7: paraphrasedistilrobertabase-v1, m8: nli-bert-large, m9: nli-distilbert-base, m10: nli-roberta-base, m11: nli-bert-large-maxpooling, m12: nli-bert-large-clspooling, m13: nli-distilbert-basemax-pooling, m14: nli-roberta-base, m15: nli-bert-base, m16: nli-bert-base-cls-pooling.

based models perform far better than lexical analysis methods. It can be observed that models based on RoBERTa, particularly roberta-large have the best performance on both minutes-transcript and minutes-minutes on the English datasets while the distilroberta-base model trained on the paraphrase dataset had the best performance on transcript-minutes and minutes-minutes on the

	Classifier	Acc	Pre	Rec	F1
Task B	Random-Forest	**0.91**	**0.71**	**0.62**	**0.66**
	SVM	0.88	0.65	0.40	0.49
Task C	Random-Forest	**0.85**	**0.42**	**0.61**	**0.5**
	SVM	0.77	0.26	0.53	0.35

Table 2: Describes the various machine learning classification approaches

Dataset	True tag	False tag	**Total**
Task B	115	731	**846**
Task C	74	660	**734**

Table 3: Class-wise distribution of Data.

Czech dataset. Models based on BERT and models trained on NLI in general performed poorly on both tasks in both languages. For lexical summarization, Doc2Vec achieves 90% and 51% for respectively. In machine learning, random forest achieves 91% and 85% and in deep learning STSB-BERT-Large outperforms all other with 94% and 81%

8 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Tirthankar Ghosal, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. Report on the SIGDial 2021 special session on summarization of dialogues and multi-party meetings (summdial). *ACM SIGIR Forum*, December 2021:1–17, 2021.

[2] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2), feb 2021.

[3] V Sowmya, K Kranthi Kiran, and Tilak Putta. Semantic textual similarity using machine learning algorithms. *International journal of current engineering and scientific research (IJCESR)*, pages 2393–8374, 2017.

[4] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 937–948, 2016.

(A) Meeting transcript segment:
(PERSON1) Yeah, that's, exactly. Ok. So Monday seminar setting. Screensharing. Yeah, yeah, so exactly. So the demo input, I think everybody I'm going through to do list. And I think that the set of the languages it should be all we have at the moment. But the main question is which should be the input language and in the call they [ORGANIZATION2] said the German would be ok for them. So I think we should be ready for for German source.
(PERSON9) Ok, nice.
(PERSON1) And like double check with them but ¡unintelligible¿ if they if they now realize that German is a bad idea then that's now a problem and we switch to English as the source.
(PERSON9) Ok. I'm going to adds this one. German as input and English as back up. Ok. Well, and translate it in all the available languages are we sure?
(PERSON1) Yeah, I would say so. Why not?It's, what would, what could be the problem?
(PERSON9) I don't know. Do we have a tested the machine translation, well the speech language translation starting from German to so all the other possible languages?
(PERSON1) So it goes via English of course. And what we have tested several times is from Czech into English and then from English into the other languages. So we have not tested-
(PERSON9) Ok.
(PERSON1) With German source, I agree. But I kind of trust the German to English [ORGANIZATION2] model. And I the the English to everything is the best that we have all the time. Like it's-
(PERSON9) Ok, ok.

(B) Meeting minutes Date: 2020/05/04
Attendees: [PERSON1], [PERSON9], [PERSON2]
Purpose of meeting: Demo preparations.
Summary of meeting:
[PERSON9], [PERSON1]
- choose [ORGANIZATION2] and [ORGANIZATION5] persons.
- From [ORGANIZATION2] is chosen [PERSON8], from [ORGANIZATION5] [PERSON8].
[PERSON9], [PERSON1]
- discuss demo input.
- German as input and English as back up.
- There are prepared some Youtube videos that are already consecutevely translated into Czech.

Cosine Distance 0.83442569631

Figure 3: An example of one of the sets in Task B where the minutes and transcript belong to the same meeting.

[5] Zongkui Zhu, Zhengqiu He, Ziyi Tang, Baohui Wang, and Wenliang Chen. A semantic similarity computing model based on siamese network for duplicate questions identification. In *CCKS Tasks*, pages 44–51, 2018.

[6] J Ramaprabha, Sayan Das, and Pronay Mukerjee. Survey on sentence similarity evaluation using deep learning. In *Journal of Physics: Conference Series*, volume 1000, page 012070. IOP Publishing, 2018.

[7] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, 2019.

[8] Kathrin Blagec, Hong Xu, Asan Agibetov, and Matthias Samwald. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. *BMC bioinformatics*, 20(1):1–10, 2019.

[9] Tao Zheng, Yimei Gao, Fei Wang, Chenhao Fan, Xingzhi Fu, Mei Li, Ya Zhang, Shaodian Zhang, and Handong Ma. Detection of medical text semantic similarity based on convolutional neural net-

(A) Meeting transcript segment:
(PERSON31) Mmm hm.
H- How bad eh, mistake was that with the deadline of ¡laugh¿ manual annotation for [PROJECT1]?
I think -
(PERSON31) Ehm -
Yeh, you aren´t really asking the right person here.I don´t know in in a sense that - After the annotations done, somebody got to analyze them.
And that person is hopefully (producer??) not me, so -
(PERSON9) Yeah. ¡laugh¿
(PERSON31) ¡laugh¿
Ehm -
And then somebody´s gonna try them up again.
That´s hopefully not me.
And that really needs to be done on time for the conference.
I think last time we had like the day before the conference.
(PERSON9) Yes, yes -
(PERSON31) So we don´t want to be any later than that.
(PERSON28) But the term - ¡parallel$_{s}peech >$
(PERSON9) But the conference is - ¡parallel$_{s}pech >$
(PERSON28) is on Saturday this Saturday.
(PERSON9) Yeah eh, and the conference is in November right?
(PERSON31) Yeah, eh in- eighteenth or so -
(PERSON9) Eighteenth, yes .
(PERSON31) So it´s is probably still time that we get the annotations done for the twentieth.
I mean we´re just just so late this year for -
(PERSON9) For everything.
So, so well, actually -
(PERSON31) ¡parallel$_{s}peech >$

(B) Meeting minutes [PROJECT1] internal Meeting
Date: 7. 9. 2020:
Attendance: [PERSON4], [PERSON13], [PERSON2],[PERSON11], [PERSON6]- Paraphrasing on
Quest:
– We should move it to “weaker” GPU (requested by PROJECT4)
– Can we do it before the end of the [PROJECT4] experiment? (“on the fly”)
- [PROJECT3]:
– in contact with [ORGANIZATION2], they provide us with their multilingual data
- [PROJECT4]:
– [PERSON12] is leaving the project!
– [ORGANIZATION4]has System Demonstration track - do we want to participate?
- [PROJECT2]:
– people from [ORGANIZATION6] (name, contact???) are also working on the decoding
constraints (or factored translation?)
- [PERSON4] getting details from [PERSON7]

Cosine Distance 0.247577452

Figure 4: An example of one of the sets in Task B where the minutes and transcript do not belong to the same task.

work. *BMC medical informatics and decision making*, 19(1):1–11, 2019.

[10] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[11] Sergio Jimenez, Julia Baquero, and Alexander Gelbukh. Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. In *In Proceedings of SemEval 2014: International Workshop on Semantic Evaluation*. Citeseer, 2014.

[12] Alice Lai and Julia Hockenmaier. Illinois-lh: A denotational and distributional approach to semantics. In *SemEval@ COLING*, pages 329–334, 2014.

[13] Johannes Bjerva, Johan Bos, Rob Van der Goot, and Malvina Nissim. The meaning factory: Formal

(A) Meeting transcript segment:
(PERSON2) So [PERSON5] is probably already there right?
That was [PERSON7], right.
So [PERSON17] would [PERSON5] join would you know?
(PERSON15) Yes yes, he he ¡unintelligible¿ minutes.
(PERSON2) Okay.
In the minutes okay.
And you are listening, you can hear us right?
Okay, so [PERSON12] can h- is listening to to the Zoom call.
But he doesnt have the microphone
Because that was causing the loop yesterday.
So its only me who has a microphone.
So [PERSON7], v-.
How much time do you need? [PERSON7] is not here.
So [PERSON7] is also trying to set up what what he posted yesterday.
So that while watching the videos, a participants who do not speak Czech should be clicking to buttons, like how well they like the current subtitles, and it would be timestamped.
And we can then align it and see like where the problems it is going to be approximate because the sync of the video.
Is not perfect, as you know, but it still, will probably be useful to to identify the the to give us some measure of the overall usability of of that.
So a please look up in your emails, because [PERSON7] sent it this morning.
The [ORGANIZATION2] document and please sign up yourself if you can to the subtitle rating uh, documents, so.
Whoever is available.
Please write your name here that is what I ’m going to highlight.

(B) Meeting minutes Test session 20200515-1000 – instead of the real demo
• Credentials:
o [URL]
o Meeting ID: [NUMBER], Password: [PASSWORD]
• Meeting is already started, the room is available until 13.00.
• ([PERSON2] will need to leave at 12.00 at the latest)
• ([PERSON12] is available only on [ORGANIZATION3], his zoom is meant only for subtitling the zoom discussion)
• Agenda:
o Summary of worker instances involved ([PERSON12], just a recap, pasting aggtable here, highlighting it)
•
• Computer names need to be shown, too, so that we can check them for load.
o [PERSON2] giving dry run of the slides.
o Czech subtitling of both Czech sample videos.
• 3min [URL]
• 15 min from this: [URL]
o Possibly English subtitling with our segmenter for English videos and our zoom discussion.
o Czech subtitling of zoom discussion ([PERSON2] and any other Czech colleague present)
o [PERSON2] giving dry run of the closing slides.

Cosine Distance 0.3254856236

Figure 5: An example of one of the sets in Task B where the minutes and transcript do belong to the same meeting but the model was not able to label it correctly.

semantics for recognizing textual entailment and determining semantic similarity. 2014.

[14] Jiang Zhao, Tian Tian Zhu, and Man Lan. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. semeval. 2014.

[15] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[16] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018, 2015.

[17] Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional

(A) Meeting transcript segment:
[ORGANIZATION2].
And that's the leadership of work packages.
Because as [ORGANIZATION7] are like not- not a standalone, not- not a full partner here, uh, they cannot lead the work package, uh, on data.
Uh, so, uh, I thought that, uh, uh, it- it would be the- good that one of the research partners, uh, does this.
And, uh, we are responsible for the multilingual MT and management and for- uh, now for minuting as well.
Uh, uh, so, uh, like we have data and minuting kind of available, uh, for [ORGANIZATION2], uh, or [ORGANIZATION11], or we can do it, uh, uh, s- as well, but, uh-
yeah, so, uh, yeah, this is-
I don't know if [ORGANIZATION11] has any opinion on this?
(PERSON14) So, uh, I think it would make more sense if, uh, [ORGANIZATION9] keeps the minuting part because you also, uh
(PERSON19) Okay.
(PERSON14) you- you've- meaning the, um, getting at the concept- you have the most involvement there.
(PERSON19) Yeah ¡unintelligible¿
(PERSON14) And I think, uh, it's okay, to have the data work package. Either us or [ORGANIZATION2] ¡unintelligible¿ [ORGANIZATION2] what they- what their position is.
(PERSON11) Yeah
(PERSON19) Yeah, so I've talked to-
yeah, [PERSON3], [PERSON3], s- please, say this- say that.
(PERSON11) So yeah, I mean, uh, ¡unintelligible¿ of course you wrote most of it, so we can also keep the minuting, otherwise since K- uh, [ORGANIZATION11] has more on- on data, I would have suggested that [ORGANIZATION11] does data and we do minuting, but I'm fine with anything else too.

(B) Meeting minutes Meeting 16. 11. 2020:
• Attendance:
– [PERSON21], [PERSON3], [PERSON22], [PERSON23], [PERSON26]
- Paraphrasing
– Deadline for student - February
– Prepare the data (provided by [PERSON16]) by the end of the month
– [PERSON23] will try translating them by the TF model
– [PERSON26] will train parent in [PROJECT1]NMT
– Questions
i. ([PERSON26]) I got model description from [PERSON16]. Still waiting for training reproduction scripts for easy training setup.
ii. We also need to evaluate the performance loss cause by compression
iii. Can we distill paraphrasing directly? By training direct paraphrasing (student) model without pivoting?
- [PROJECT5]
– [PERSON23] is processing the segments provided by [ORGANIZATION7]
– waiting for [ORGANIZATION7] ([PERSON18])
– trying to submit paper to [ORGANIZATION1]
- [PROJECT2] (data augmentation)
– aiming for a workshop rather than [ORGANIZATION1]
- [PROJECT6]
– [PERSON22] is training an already constrained model
— comparing it with constrained decoding on inflected languages
– Next step is focus on constraining in inflected languages
- Implementation

Cosine Distance 0.77585155236

Figure 6: An example of one of the sets in Task B where the minutes and transcript do not belong to the same meeting but the model was not able to label it correctly.

deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382, 2015.

[18] Basant Agarwal, Heri Ramampiaro, Helge Langseth, and Massimiliano Ruocco. A deep network model for paraphrase detection in short text messages. *Information Processing & Management*, 54(6):922–937, 2018.

[19] Rafael Ferreira, George DC Cavalcanti, Fred Freitas, Rafael Dueire Lins, Steven J Simske, and Marcelo Riss. Combining sentence similarities measures to identify paraphrases. *Computer Speech & Language*, 47:59–73, 2018.

[20] Yushi Homma, Stuart Sy, and Christopher Yeh. Detecting duplicate questions with deep learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS*, 2016.

[21] Jiangping Huang, Shuxin Yao, Chen Lyu, and Donghong Ji. Multi-granularity neural sentence model for measuring short text similarity. In *International Conference on Database Systems for Advanced Applications*, pages 439–455. Springer, 2017.

[22] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. Semantic text matching for long-form documents. In *The World Wide Web Conference*, pages 795–806, 2019.

[23] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[24] Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. Sentence similarity learning by lexical decomposition and composition. *arXiv preprint arXiv:1602.07019*, 2016.

[25] Wenpeng Yin and Hinrich Schütze. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911, 2015.

[26] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272, 2016.

[27] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52, 2010.

[28] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[29] Jiahui Liu and Larry Birnbaum. Measuring semantic similarity between named entities by searching the web directory. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 461–465, 2007.

[30] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606, 2016.

[31] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.

[32] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326, 2015.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[36] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.

An End-to-End Multilingual System for Automatic Minuting of Multi-Party Dialogues

Aakash Bhatnagar[&], **Nidhir Bhavsar**[$], **Muskaan Singh**[#] and **Petr Motlicek**[#]
[&] Boston University, Boston, Massachusetts
[$]University of Potsdam, Potsdam, Germany
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
aakash07@bu.edu, bhavsar@uni-potsdam.de,
(msingh,petr.motlicek)@idiap.ch

Abstract

In this paper, we present a pipeline for automatic minuting. This pipeline is an end-to-end system for minuting the multiparty dialogues of meetings. It provides multilingual communication and collaboration, with a specific focus on Natural Language Processing (NLP) technologies: Automatic Speech Recognition (ASR), Machine Translation (MT), Automatic Minuting (AM), Topic Modelling (TM), and Named Entity Recognition (NER). Our summarization model achieves a ROUGE-1 score of 0.45, a BLEU score of 7.069, and a BERT score of 0.673. Our translation model also achieves a high average BERT score of 0.848 across five different languages (de,fr, en, it, and hi). We make our code available at https://github.com/aakash0017/Paclic-summarization-pipeline

1 Introduction

Since the COVID-19 epidemic, a sizeable portion of the working population—particularly those employed in the information technology (IT) sector and academia—has expanded dramatically in virtual meetings. Meetings are, without a doubt, the most important element in fostering teamwork and effective back-and-forth communication. There are numerous Natural Language Processing (NLP) technologies available that give users a complete online interaction experience. The interpretation of these online interactions during remote conferences or meetings is crucial in the globally interconnected world of today.

Summarizing meetings in the form of structured minutes from speech and it can potentially save up to 80% of time.

We realized that generating meeting minutes is a task that is still performed manually and requires a lot of time. Through this paper, we try to solve three major problems encountered during multi-party dialogues via our proposed system.

First, as mentioned above, manually generating minutes consumes much time. Each annotator has to go through hours of recording before writing minutes. Also, each annotator may have a different vocabulary and style, leading to inconsistency in the meeting-minutes format. We try to solve this problem by generating meeting minutes in a consistent format for multi-party conversation(MPC). We incorporate large pre-trained transformer models fine-tuned on MPC meeting datasets.

Second, as globalization is increasing, companies have offices worldwide. Hence, to overcome the language barrier, there is a need for a system that can provide translation of meeting transcripts, meeting minutes, and meeting topics. We provide quick and straightforward translation in five different languages: French(fr), German(de), Russian(ru), Italian(it), and Hindi(hi), allowing businesses to save time and enhance productivity. To further optimize the translation process, we provide isometric translation [1], which generates outputs of length similar to the source length. We believe that isometric translation is the next leap towards a more synchronous auto dubbing process, which can enhance the meeting experience of non-English speaking users.

Lastly, the enormous increase in online meetings and conversations has led to a massive stack of unordered data. It can be cumbersome for users to select the appropriate meeting for their needs. We try to differentiate the generated minutes into multiple segments and align them with corresponding topics derived accordingly. This provides a gist of the meeting without the user listening to the whole recording.

The paper is organized as: Section: 2 brief about the existing work in text and meeting summarization. The proposed methodology is described in Section: 3, where we employ a 3 stage pipeline: ASR; automatic speech recognition, multi-party meeting summarization, isometric translation, and topic segmentation. The experimental setup with the dataset, hyper-parameter settings, and training are in Section:4 with their corresponding results and in Section: 5. Finally, the paper concludes with future prospects in Section: 6.

2 Related Work

In this section, we describe the existing work on text summarization in Section 2.1 and meeting summarization in Section 2.2.

2.1 Text Summarization

Majority of the prior work on meeting summarization investigates how to generate better summaries for news/media article data, such as CNN/Daily Mail [2], Newsroom [3], etc. other tries to summaries scientific documents, such as SciSumm Corpus [4]. However, our paper mainly focuses on meeting summarization which is comparatively a more challenging task. However we do tend to infer some attributes of normal text summarization, which includes both extractive & abstractive methods. Moreover, the topic of meeting summarization, especially automatic minuting has been a demanding research problem across the community [5] [6] [7] and has become a huge part of the text summarization area.

2.2 Meeting Summarization

Since the advent of COVID-19 and the majority of work shifting online, there has been a lot of interest gathering around multi-party dialogue summarization. However, the fundamental idea behind meeting recording summaries has existed for quite some time. [8] suggested a extractive meeting summarization approach using graphs constructed on topical/lexical relations. However, a study conducted by [9] stated the difference between meeting summarization of multi-party transcripts over the Natural Language Generation models for generating unfocused summaries. They proposed multi-modal hierarchical attention across three levels: segment, utterance, and word and suggested a joint model of topic segmentation and summarization. [10]. Next, [10] attempted to pre-train MPC-BERT to find the inherent complicated structure in MPC via crucial interlocutor and utterances. [11] proposes a novel abstractive summary network that adapts to the meeting scenario. It follows a hierarchical structure to accommodate long meeting transcripts and a role vector to depict the difference among speakers.

The aforementioned work, however, creates meeting summaries, whereas our suggested approach attempts to create meeting minutes from the ASR generated transcripts.

3 Proposed Methodology

We propose a pipeline that utilizes a speech-to-text transcription service and a meeting summarization module. Additionally, we provide functionality of topic extraction and isometric translation (German(*de*), French(*fr*), Italian(*it*), Russian(*ru*), and Hindi(*hi*). As depicted in the figure 1, the system accepts a {*.mp3, .mp4*} file consisting of multi-party conversations in English(*en*). Next, we generate ASR output from the input files, which are then utilized by our system to generate meeting minutes. From subsections 3.1 through 3.4, we provide a details overview of various components of our proposed architecture.

3.1 Automatic Speech Recognition (ASR)

For generating optimal transcripts, we use Amazon Transcribe[1], which is a Speech-to-text service offered by Amazon AWS. It holds the largest share in the cloud computing market. Their current English speech recognition model has achieved a word-error-rate (*WER*) of 6.2%. To convert meeting recordings to transcripts, the data must first be uploaded to the Amazon Simple-Storage-Service (*Amazon S3*), which is then used by the Amazon transcribe Speech-to-text API to generate time-sequence order transcripts, with both speaker and utterances stated separately. To handle this, we define a post-processing function that align speaker roles with corresponding utterances, as shown in figure 1. Our system accepts a number of speakers as an argument before applying ASR transcriptions. However, the argument is set to 2 and accepts a maximum value of 10.

3.2 Meeting Summarization

The meeting summarization module generates meeting-minutes from the processed transcripts.

[1] https://aws.amazon.com/transcribe/

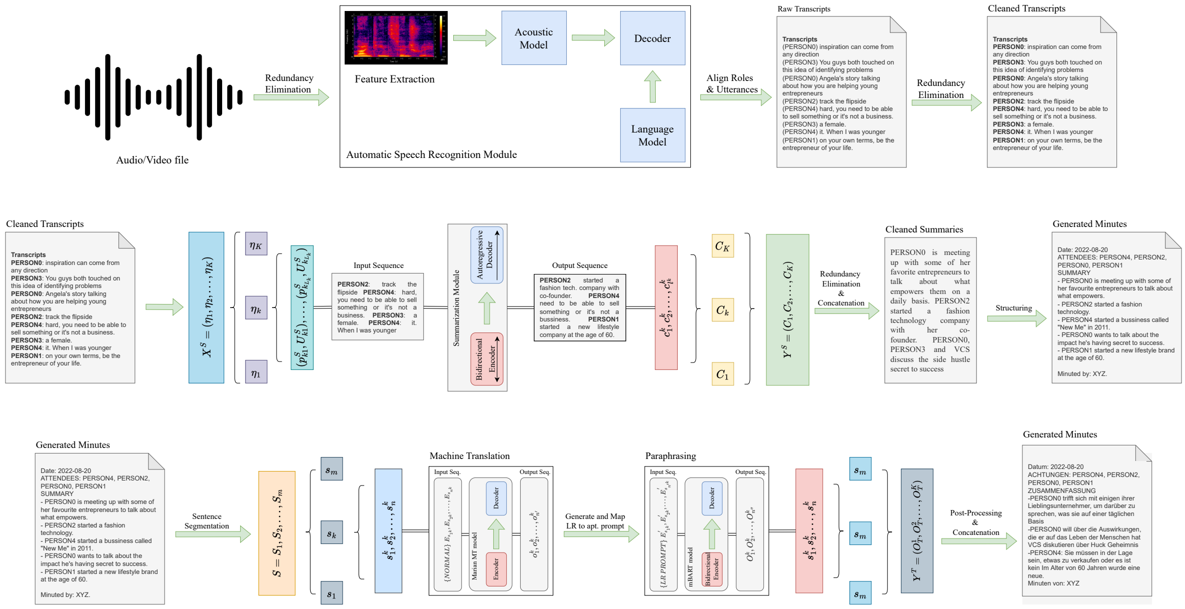


Figure 1: Displays the entire architecture of our proposed systems.

The task of automatic-minuting differs distinctively from the summarization task. Minuting is primarily concerned with capturing and providing a third-person perspective of important points raised throughout the meeting, whereas summarizing is more concerned with delivering a piece of concise information and not reflecting small details. Our pipeline overcomes one major drawback of manual minuting; that the minutes format and language vary across different annotators.

The meeting summarization module is divided into three main parts. First, we start by preprocessing input transcripts, apply redundancy elimination and segmentation. Next, we apply our meeting summarization model. Finally, we filter the output using an unsupervised redundancy elimination method to obtain the processed minutes.

The majority of dialogue summarization system lacks the ability to refrain from redundancies. Besides that they are also limited to a specific length of input sequences for accurate text generation. Our proposed system tries to tackle these issues using the Redundancy Elimination and Segmentation module. We employ some handcrafted rules and pre-processing techniques to process the input utterances obtained from previously generated transcripts. First, text cleaning procedures are used to get rid of any repeats, pauses, and interruptions in the text. These utterances are then filtered using a custom stopwords we define from publicly available meeting summarization corpuses like AMI [12] & ISCI [11]. Next, we utilize some brute force approach to slice the non-redundant transcripts to address the limitation of length constraints of input sequences. Currently our system support segmentation for varying token lengths of *512*, *768* and *1024* respectively.

We use a finetuned BART-large model [13] [2] for our primary summarization task. BART is a denoising autoencoder for pretraining sequence-to-sequence models. The model is trained by using arbitrary denoising functions to distort text and then instructing it to recreate the original content. Using BART provides the ability to use bi-directional attributes when operating on sequence generation tasks which makes it useful for abstractive text summarization. While BERT cannot adopt a bidirectional mechanism for sequence generation, BART exploits the GPT-2 architecture for predicting the following words with the help of words encountered previously in the current sequence. Hence, we primarily test the pipeline with various BART-based setups. However, we majorly experiment with a fine-tuned version of BART trained simultaneously on XSum [14] & SAMSum[15] datasets.

The generated summaries contain a sufficient amount of information, although they are not entirely adequate. There might be an inclusion of casual discussion or other unnecessary information. This problem is addressed with TextRank. Based on our experimentations, we found out that from

[2] https://huggingface.co/lidiya/bart-large-xsum-samsum

the whole report, the model typically catches 15% of trivial and unnecessary information. We rank the summary lines in increasing order of their importance and exclude out bottom 15% of the lines to obtain a "gold span" of the summary. To further compress the summaries, we add appropriate pronouns, eliminate grammatical inconsistencies wherever possible, and filter the final chain of conversation threads by excluding unnecessary words using stopwords set that we internally develop by observing the generated summaries.

3.3 Isometric Machine Translation

The Machine translation module provides set up the capability to generate transcripts, minutes, and topics in five different languages *de, fr, it, ru, hi*. For all these languages, we provide a user with isometric translation output. Isometric MT is the concept of generating translation that falls within the source length range of $\pm 10\%$. This feature helps to generate synchronous outputs upon text-to-speech conversion. For implementing isometric translation, we develop a multitask learning model similar to [16]. We use fine-tuned OPUS-MT [17] model for translation and fine-tuned mBART [18] for paraphrasing. However, our isometric translation module works best for French, German, and Russian languages as we implement a paraphrasing model to enhance the vocabulary. Hindi and Italian translation does not contain a paraphrasing model, but the use of prompt engineering techniques enables them to achieve a high BLEU score and BERT score.

3.4 Topics Modelling

DeepCon also provides a feature for automatic topic extraction based on Named Entity Recognition that extracts the top-k repeating n-grams from the transcripts. We use Yake[3] library for extracting named entities. Our system can also translate the keywords into the five different languages *de, fr, it, ru, hi*. We intend to generate these topics or keywords in order to provide a comprehensive abstract view of meeting discussions with the generated minutes.

4 Experimental Setup

In this section, we describe dataset details in section 4.1, hyper-parameter setting in section 4.2 and training procedures in section 4.3

4.1 Dataset

As stated, our summarization module utilizes a BART model fine-tuned on both XSum and SAMSum datasets. XSum dataset includes short summaries of articles and discussions, whereas SAMSum is a multi-party meeting conversation dataset usually comprising casual and friendly conversations. Training model on these two datasets allows it to grasp summarization both at the syntactic and morphological levels. For evaluating our proposed summarization models, we used the publicly available ELITR Minuting Corpus [19]. The corpus is divided into 3 subtasks. However, we used the Task-A dataset with the dataset distribution of $85, 10$, and 25 instances for train, validation, and test set, respectively. Each instance comprises i) a meeting transcript and ii) one or more than one human annotated meeting minutes. Table 1 shows the statistics of the dataset that we have used in experimentation.

Next, for the isometric translation module, we experimented with multiple datasets across previously specified languages for both the machine translation and paraphrasing modules. For machine translation, we majorly use the Multilingual Speech Translation Corpus (MuST-C) [20]. We also utilise the Statistical Machine Translation Dataset (WMT) [21] for German (de) & Russian (ru) text inputs. Additionally for translating Hindi (hi) we use the IIT-B Hindi-English Corpus [22]. Next, we use a combination of Opusparcus [23] and PAWS-X [24] datasets for most of our Paraphrase training tasks, However, due to unavailability of PAWS-X dataset for Russian (ru), we utilize the Tapaco [25] dataset which is a sub-extracted paraphrase corpus derived from the Tatoeba database [26]

4.2 Hyper-parameter Settings

We used 4 Tesla V100-PCIE GPUs for all experiments with a memory size of 32510 MiB each. Due to resource constraints, we train each of our models both for summarization and isometric machine translation for 1 epoch, each with a batch size of 32. Our fine-tuned models are trained on a learning rate configuration of 2×10^{-5}. For finetuning the underlying summarization model, we use the following configurations: *'max input length' = 512*, *min target length = 128*. Next, for the isometric machine translation module for both our machine translation and paraphrasing model training, we implement the *AdaFactor* optimizer, which inter-

[3] https://github.com/LIAAD/yake

Table 1: Represent the various statistics calculated on both the SAMSum and ELITR datasets. This includes the No. of dialogues, No. of turns, No. of speakers, No. of average turn lengths, Length of dialogues, Summary lengths, and Percentage of compression.

Datasets	# diag.	# turns	# speakers	avg. turn len.	# len. of diag.	# summary len.	% comp.
SAMSum	16.4K	11.2	2.4	9.1	124	23.4	82.12
ELITR	124	254.4	5.8	9.7	8890.8	387	95.65

nally adjusts the learning rate based on the scale parameter and relative/warmup steps.

4.3 Training

In this section, we discuss all experiments performed for our proposed system. We experimented with various automatic speech recognition (ASR) models for generating MPC transcripts. This includes Wav2Vec[4] [27] model trained on the MInDS-14 [28] dataset. We used the Word-Error-Rate (WER) to evaluate these models. However, most of our trained models generated the speech-to-text output with reasonably high WER scores, runtime, and Samples per Second during testing. Additionally, the transcripts that were generated appear cluttered with extra incomplete content. Thus finally, we decided to use the Amazon Transcribe service to generate meeting transcriptions for further processing.

Next, we experiment with multiple summarization models using T5 [29], Pegasus [30], RoBERTa2RoBERTa [31], distilBART [32], etc. However, the BART-based pipeline performed better than the rest. Table 2 represent the scores evaluated on the ELITR Task-A test dataset. Our experiments include fine-tuning these pre-trained models on various summarization datasets. This includes, CNN/DailyMail, XSUM, SAMSUM and AMI Corpus.

We also implement a singleton MT5 model that performs translation and paraphrasing of 5 supported languages using the prompt engineering method. In this approach, we use two additional prompts combined with length prompts: 1) Translation and 2) Paraphrasing. The translation prompt signifies that the model will translate the given input, and the paraphrasing prompt signifies that the model will generate isometric sentences from the translated sentences. We use the MUST-C dataset for translation and PAWS-X and Topaco for paraphrasing. However, the sentences generated by this MT5 model are very redundant and non-contextual.

Table 2: Performance of different baseline models considered during experimentation. This includes Rouge-1, Rouge-WE, BLEU score, BERT-F1 score, TF-IDF score.

Models	R1	RWE	BLEU	BERT	TFIDF
BART	0.297	0.162	2.907	0.563	0.19
DistilBART	0.375	0.205	6.535	0.620	0.25
T5	0.406	0.229	6.278	0.615	0.31
Ours	**0.45**	**0.298**	**7.068**	**0.673**	**0.38**

Next, we adopt a prompt-based few-shot learning strategy for the paraphrasing task. The model utilizes a small sample of the training dataset(approx 500) and then tries to integrate the derived model with the predictions obtained from the MT model. The comparative scores achieved by assessing using the same technique are listed in table **??**. The few-shot model can adequately constrain the output length while preserving the MT's semantical aspects.

5 Results and Analysis

As said earlier, our system accepts total speakers in the range $\{2, 10\}$. Also, providing an exact count of total speaker value shows that it helps the Amazon Transcribe model to generate the best results and align each speaker utterance with its audio counterpart.

We evaluate our proposed summarization model based on the following metrics. This includes i) ROUGE-N; to match n-grams between system predictions and target gold spans. ii) ROUGE-WE; Since ROUGE-N is extremely biased toward lexical similarities, we attempt to compare the projected summaries using the word embeddings based ROUGE as described in [33]. iii) BLEU; though preferred for evaluating machine translation output, we use this metric to calculate the quality of generated summaries. iv) BERT-score; it calculates the semantic relatedness by aligning the sentence representation of both reference and hypothesis using the cosine similarity. v) TF-IDF; works by calculating the importance of each word/token in generated output based on its occurrence in the doc-

[4] https://huggingface.co/facebook/wav2vec2-base-960h

language	model	dataset	BLEU Score	BERT Score	Length Ratio	Length Range
de	OPUS-MT	MuST-C + WMT	**42.3**	**0.85**	**1.087**	49.81
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	29.1	0.83	1.04	50.55
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	29.9	0.83	1.05	**51.95**
it	OPUS-MT	MuST-C	**34**	**0.84**	**1.045**	**57.032**
fr	OPUS-MT	MuST-C	**44.8**	**0.87**	1.08	49.6
	OPUS-MT+ MT5	MuST-C	42.3	0.85	1.12	51.3
	OPUS-MT + MT5	MuST-C	38	0.86	1.11	46.4
	OPUS-MT + few short mBART	MuST-C + Opusparcus + Paws-X	40.9	0.85	**1.03**	57.33
	OPUS-MT + mBART	MuST-C + Opusparcus + Paws-X	41.2	0.85	1.04	**61.81**
ru	OPUS-MT	MuST-C + WMT	**22.7**	**0.84**	**1.005**	54.517
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	20.8	0.82	0.95	58.934
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	21.7	0.83	0.967	**62.475**
	MT5	MuST-C + WMT	5.6	0.76	0.732	19.3
hi	OPUS-MT	IITB-En-hi	**11.9**	**0.84**	**0.941**	**42.521**

Table 3: Evaluation scores of various experiments. In this table, we state the language-wise experiments along with the datasets used

ument. Table 2 shows performance analysis of our proposed summarization models and its comparison to various other summarization models. As is evident, our suggested approach produces better results when compared to the other models, and by a significant margin. This indicates that our generated meeting minutes are more accurate regarding Grammatical Correctness and Fluency than the other recent summarization models.

We use BLEU, BERT-score and length compliance metrics to evaluate isometric translation outputs. As mentioned earlier, a statistical method evaluates based on n-grams in translated and reference text and rates the quality of the predictions. BERT-score and Length Compliance metrics are specially designed for the task of isometric MT. The Length Compliance metrics comprise 2 measures, a) length ratio, calculated by matching the length of predicted text against the gold-span targets, and b) length range, which measures the percentage of sentences that falls within the ±10 ideal span of the length-ratio. Table 3 represents the evaluation scores obtained by training various models during experimentations across the previously mentioned MuST-C and IIT-B test datasets. First, the best-performing models for the isometric task for each source language have a lower BLEU score. This suggests that the isometric constraints can affect the calculated BLEU score since it is character-dependent. However, emulating the predicted MT text via the paraphrase model suggests a higher BLEU score. This is because the paraphrasing module modulates the sentence length to conform to the interchangeable vocabulary.

6 Conclusion

The proposed pipeline efficiently handles audio/video files and generates meeting minutes, translations and topics. However, the pipeline does not extract any feature from the video. We believe that using video frames and fusing them with the embeddings of ASR output can generate some quality results. By Introducing the multi-modality aspect, we can further leverage the essential information the video provides. The task of multi-modal fusion poses a significant challenge, and thus we hope to counter it in our upcoming projects. This pipeline can also be extended as an API service for developers to incorporate Auto-minuting functionality in their systems.

7 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *ACL*, 2022.

[2] Ramesh Nallapati, Bing Xiang, and Bowen Zhou. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023, 2016.

[3] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *CoRR*, abs/1804.11283, 2018.

[4] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*, 2019.

[5] Lu Wang and Claire Cardie. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A., June 2014. Association for Computational Linguistics.

[7] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Yun-Nung Chen and Florian Metze. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 377–381, Montréal, Canada, June 2012. Association for Computational Linguistics.

[9] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. *CoRR*, abs/2106.01541, 2021.

[11] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. End-to-end abstractive summarization for meetings. *CoRR*, abs/2004.02016, 2020.

[12] Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *CoRR*, abs/1809.05715, 2018.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[14] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745, 2018.

[15] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237, 2019.

[16] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 379–385, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[17] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020.

[19] Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondrej Bojar. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. 2022.

[20] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[22] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT bombay english-hindi parallel corpus. *CoRR*, abs/1710.02855, 2017.

[23] Mathias Creutz. Open subtitles paraphrase corpus for six languages. *CoRR*, abs/1809.06142, 2018.

[24] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828, 2019.

[25] Yves Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association.

[26] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, 2018.

[27] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[28] Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal Lis, Eshan Singhal, Nikola Mrksic, Tsung-Hsien Wen, and Ivan Vulic. Multilingual and cross-lingual intent detection from spoken data. *CoRR*, abs/2104.08524, 2021.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[30] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777, 2019.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[32] Sam Shleifer and Alexander M. Rush. Pre-trained summarization distillation. *CoRR*, abs/2010.13002, 2020.

[33] Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Law Retrieval with Supervised Contrastive Learning Using the Hierarchical Structure of Law

Jungmin Choi[1,2], **Ukyo Honda**[1,2], **Taro Watanabe**[1], **Kentaro Inui**[2,3], **Hiroki Ouchi**[1,2]
[1] Nara Institute of Science and Technology
[2] RIKEN
[3] Tohoku University
choi.jungmin.ce6@is.naist.jp, honda.ukyo.hn6@is.naist.jp,
taro@is.naist.jp, inui@ecei.tohoku.ac.jp, hiroki.ouchi@is.naist.jp

Abstract

We study the information retrieval task to identify the relevant law articles for a query on a legal issue in when the legal system in question is statute law. In recent years, the mainstream approach has been to calculate the similarity between the query and each article using pre-trained language models. However, such methods have a weakness in retrieving relevant articles that have low n-gram similarity scores with the query. In this work, we show that in such hard cases, the articles tend to be of the same class as articles with high n-gram similarity scores in the hierarchical structure of statute law, for instance, the Japanese Civil Code. From this observation, we hypothesize that by making articles of same class close to each other in the feature space, we could make it easier to retrieve the above mentioned hard articles. Our proposed method realizes this by supervised contrastive learning using the hierarchical structure. Experimental results show that the proposed method achieves higher performance in retrieving the correct articles with low n-gram similarity to the query.

1 Introduction

Law is one of the domains where application of natural language processing is expected to bring immense benefit to the society. According to a survey conducted by Japan Federation of Bar Associations[1], nearly half of those who visited law firms or legal support centers for consultation answered that they had hesitated consulting lawyers before visiting. Some of the most frequently cited reasons include unapproacheable image of lawyers, anticipated difficulty communicating with them, and concerns about whether the issues will be taken seriously. This indicates that people often have psychological obstacles to accessing legal services. It is an important societal task to lower the barrier to accessing legal services and facilitate function of law throughout the society.

A solution to this problem by natural language processing is to build an information retrieval system that suggests relevant laws to user-given queries regarding their legal issues. It will provide the user with an approximate idea about how their issues could be described in legal terms, which will enable them to further search for more refined information without necessarily having to consult legal professionals or ask better informed questions when they seek legal support.

When developing such a system, it is essential to take into account the characteristics of the legal system of interest. Depending on the legal system, laws are written in vastly different style and structure. In this regard, legal systems can be categorized into two broad categories, case law and statute law. Case law, which includes the legal systems of the United Kingdom, United States, Canada, etc., is law that is based on past judicial decisions, while statute law, examples of which are the legal systems of Germany, France, Japan, etc., is written law passed by a body of legislature. With case law, the task in question would be to retrieve relevant cases, and whereas

[1] https://www.nichibenren.or.jp/library/ja/jfba_info/publication/data/shimin_needs.pdf

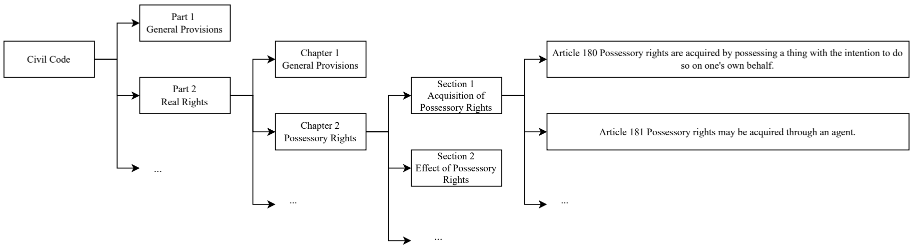


Figure 1: Hierarchical structure of Japanese Civil Code.

with statute law, it would be to retrieve relevant law articles. Note that a case document is written on actual legal disputes and therefore concrete in nature, and a law article is written in abstract legal terms and are organized in hierarchical structure.

In this paper, we focus on statute law. Our task here is to retrieve relevant articles to a given query from a set of candidate articles. As mentioned above, the language in which statute law is written is significantly different from ordinary language. For our purpose, which is to assist users who might not be familiar with those abstract legal terms, it is particularly important that our system can correctly retrieve relevant articles when the query has little to no overlap in terms of vocabulary with the relevant articles.

In this task, the mainstream approach has been to employ pretrained language models (Wehnert et al., 2021; Nguyen et al., 2021; Shao et al., 2021). However, it has been pointed out that these approaches perform poorly when detecting relevance between the query and article is semantically involved (Rabelo et al., 2021).

To address this problem we propose a method to use the hierarchical structure of statute law, which is not fully utilized in previous work. Statute law is typically organized in hierarchical structure. For example, articles in the Japanese civil code are classified on five levels: part, chapter, section, subsection, and division. See Figure 1. At each level, articles in the same class share the same topic. Within the framework of retrieving articles according to their similarity to the query in the embedding space, we hypothesize that by training the model to map articles in a same class closer in the embedding space and articles in different classes apart, we would be able to leverage the structure information and obtain better embeddings. This point will be elaborated in Section 5, Supervised Contrastive Learning for Law Retrieval.

We have conducted experiments using as the benchmark the data set created for Task 3 of the workshop, Competition on Legal Information Extraction/Entailment (COLIEE), where participants compete building systems that automatically answer Japanese bar exam problems. The experimental results show that our approach outperforms previous approaches which supports the effectiveness of contrastive learning as a way to incorporate hierarchical structure in the embeddings.

2 Related Work

2.1 Document Retrieval

The task of retrieving relevant documents to queries has been a central component of information extraction and question answering (Narasimhan et al., 2016; Kwok et al., 2001; Voorhees, 2001). While early research on this task has focused on sparse bag-of-words representations, recent advances in computational resources has inspired a plethora of research using neural network (Mitra and Craswell, 2017).

In general, neural models for document retrieval use vector representations of text, and contain a large number of parameters to be tuned. Therefore, they typically require a large training data set.

To mitigate the computational burden, a body of recent work has adopted a two-stage retrieval and ranking pipeline. At the first stage, they retrieve a large number of documents using sparse high dimensional query/document representations, and at the second stage, they rerank the documents with learned neural models (Nogueira and Cho, 2019; Yang et al., 2019). While this approach has achieved state-of-the-art results on information retrieval benchmarks, it suffers from the upper bound imposed by any recall errors in the first-stage retrieval model (Luan et al., 2020).

A strong alternative is to perform first-stage retrieval using learned dense low-dimensional encodings of queries and documents. Reimers and Gurevych (2019) has shown tha their dual encoder model which scores each document by the inner product between its encoding and that of the query perform well and efficiently. Karpukhin et al. (2020) outperformed traditional sparse vector space models such as TF-IDF and BM25 for retrieving answers from open-domain context by a simple dual-encoder framework.

2.2 Legal Information Retrieval

2.2.1 Case Retrieval

Searching through a large collection of previous cases (court decisions) for ones that apply to a particular situation is an important part of day-to-day work of legal professionals. Hence, there have been efforts to automate this task in jurisdictions from all over the world (Hafner, 1980; Parikh et al., 2021; Xiao et al., 2019; Rabelo et al., 2021; Chalkidis et al., 2020).

While this task can be formulated as a special case of document retrieval, it has been noted that document retrieval methods that perform well with general data sets do not transfer easily to the legal domain, for reasons such as the large number of candidate documents, verboseness of each case documents, the definition of relevance in the legal scenario being beyond the general definition of topical relevance (Shao et al., 2020; Alberts et al., 2021; Van Opijnen and Santos, 2017). Ma et al. (2021) has applied traditional language model and showed that it outperformed neural models. Rosa et al. (2021) has shown that their method of splitting case documents into segments and applying BM25 to rank cases by similarity to the query perform competitively against neural models.

2.2.2 Statute Retrieval

In recent years, the mainstream approach has been to use TF-IDF and pretrained language models.

Wehnert et al. (2021) computes the cosine similarity of each query-article pair based on Sentence-BERT (Reimers and Gurevych, 2019) representation and TF-IDF, sum the two cosine similarity values, and classify the pair as relevant if the value exceeds the predetermined threshold. If none of the articles has similarity higher than the threshold, the article with the highest similarity is selected as the relevant article. The threshold is determined so that the score will be highest if applied to the validation set. They use the English version, and employ a Sentence-BERT model pretrained with millions of paraphrase pairs. The problem with this approach is that it is hard to grasp the similarity between a query and article when it requires high-level semantic matching, e.g., when the query involves concrete examples of abstract concepts in the relevant article.

Nguyen et al. (2021) treats this task as a binary classification problem. They concatenate the query and the article with a SEP token to make a single sequence, applies linear transformation to the BERT features corresponding to the CLS token of this sequence and conduct binary classification whether or not the query and article are relevant. In order to mitigate the label imbalance problem, i.e., irrelevant query-article pairs far outnumbering relevant ones, they only use pairs whose article places in the top 150 among all articles in terms of TF-IDF similarity with the query. They use the Japanese version. Like Wehnert et al. (2021), it has difficulty in semantic matching. It also suffers from a strict upper bound imposed by any recall errors in the first stage where they limit the candidate to top 150.

2.3 Supervised Contrastive Learning

Contrastive learning is a method which aims to obtain effective representation of objects by gathering semantically similar ones in closer proximity in the embedding space and distancing dissimilar ones. Khosla et al. (2021) has introduced supervised contrastive learning for image classification task where they trained the model so that images that belong to same classes will have closer embeddings and those with different classes distant embeddings. Gao et al. (2021) applies such framework to sentence representations, and proposes simple contrastive learning method with a supervised setting. Using natural language inference (NLI) data sets, the method learns

Table 1: Number of Queries by Number of Relevant Articles

Number of Relevant Articles	1	2	3	4	5	6	Total
Number of Queries	567	125	25	5	2	1	725

to bring sentences with entailment relations closer together as positive example pairs and to move sentences with contradiction relations away from each other as negative example pairs. The method significantly outperforms previous methods on a variety of semantic textual similarity tasks.

3 Task

3.1 Overview

We focus on COLIEE Task 3 as our benchmark for the legal information retrieval task. In the data set, query statements regarding legal issues are each paired with sets of relevant law articles. The query statements mostly describe specific situations and tends to be written with ordinary vocabulary, while articles are written using abstract, legal terms, which makes it an appropriate benchmark for a system that identifies relevant articles to queries written by non-experts as described above.

The data consists of train and test data, and there the original version written in Japanese and a translated version in English. Every year, the latest bar examination problems become the test data, and the past problems, including the previous test data, form the train data. For the competition in 2021, the training data consisted of 725 examples of queries and corresponding sets of relevant articles, while the test data had 81 such examples. Table 1 shows the distribution of number of relevant articles to each query. Figure 2 is an example from the data.

3.2 Definition

Formally, the task is defined as follows. There is a set of N_A Japanese Civil Code articles $\mathbb{A} = \{a_1, a_2, \dots a_{N_A}\}$ and a set of N_Q queries $\mathbb{Q} = \{q_1, q_2, \dots, q_{N_Q}\}$. $\mathcal{D} = \{(q_i, \mathbb{A}_i)\}_{i=1}^{N_Q}$ is a set of pairs where each query q_i is paired with the set of its relevant articles $\mathbb{A}_i \subset \mathbb{A}$. Given a query q_i, the task is to find $\mathbb{A}_i$. It is assumed that for each q_i, exactly one such non-empty subset of $\mathbb{A}$ exists.

Query In cases where an individual rescues another person from getting hit by a car by pushing that person out of the way, causing the person's luxury kimono to get dirty, the rescuer does not have to compensate damages for the kimono.

Relevant Article Article 698 If a manager engages in benevolent intervention in another's business in order to allow a principal to escape imminent danger to the principal's person, reputation, or property, the manager is not liable to compensate for damage resulting from this unless the manager has acted in bad faith or with gross negligence..

Figure 2: An example of Task 3. Note that here, ”getting hit by a car” is a concrete example of ”imminent danger” and ”kimono to get dirty” is that of ”damage resulting from this”

Table 2: R@10 scores of relevance prediction to query-article pairs by Wehnert et al. (2021). All represents the whole validation set, Easy represents Easy is the subset of All with higher n-gram similarity, Hard represents the subset with lower n-gram similarity. We show the recall score for each set.

Easy	Hard	All
97.30	42.67	60.71

4 Problems and Analysis of Previous Work

4.1 Problems

In previous studies, when the n-gram similarity between the query and the relevant article is low, it is difficult to classify the article as correct (Rabelo et al., 2021). This is the case, for example, when the query is a description of concrete facts, and it is necessary to correspond the norms of the article to the facts in order to determine the relevance of the relevant article to the query.

4.2 Analysis

In order to confirm this problem statistically, the validation data set was divided according to n-gram similarity, and the scores of each division is shown below. In the validation data set, there are a total of 112 pairs of query and relevant article. We apply TF-IDF vectorization to each query and article, and compute the cosine similarity of these vectors.

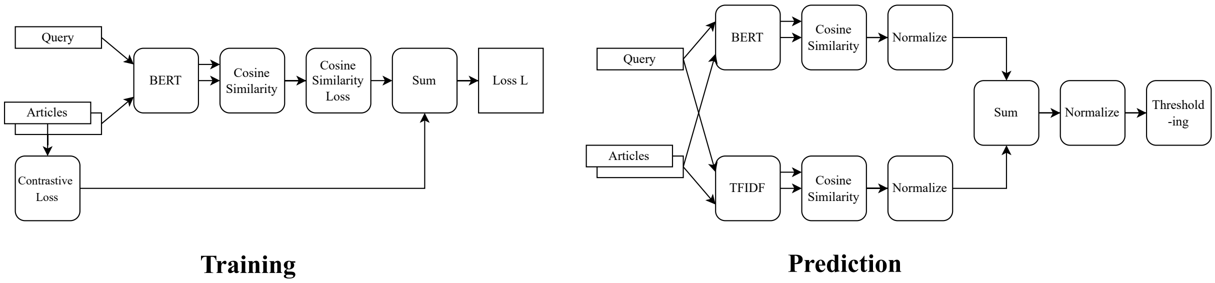


Figure 3: Overview of our proposed method.

There are 75 pairs where the normalized cosine similarity[2] between the query and the relevant article is less than 1. We call such cases Hard, and other cases Easy. The entire cases are called All. Table 2 show the accuracy score for each division from the model that replicates (Wehnert et al., 2021). Note that it is particularly low for Hard.

5 Supervised Contrastive Learning for Law Retrieval

Analyzing the 75 query-article pairs (q_i, a_j) with cosine similarity less than 1, we have found that there are 26 of cases where the article a_k with the highest cosine similarity to the query q_i and the correct article a_j are in the same section. This suggests that even a correct article a_j with low n-gram similarity can be pulled up to the top of the search results if the expressions of articles belonging to the same sections are made closer to each other. Based on this hypothesis, we propose contrastive learning using the section information of articles as labels. The proposed method uses the section information because almost every article belongs to some section and each section has on average approximately 11.3 articles, which makes it an appropriate way to divide articles into semantically similar groups.

Based on this analysis, we expect that it may be easier to obtain relevant articles with low n-gram similarity to the query by putting articles belonging to the same section closer to each other in the embedding space. Therefore, we propose a contrast learning method using sections as labels.

Previous methods have included hierarchical information (e.g., Part 1, Chapter2, Section3) in the input sentences, however, it is unlikely that these methods have successfully incorporated the hierarchical structure. Intuitively, simply including the hierarchical information in the input will not directly incentivize the model to pull same-section peers together, and might even hurt learning by forcing some tokens at the end of the input to be cut off because of sequence length limit of the model. Indeed, in our preliminary experiments, we found no advantage of including hierarchical information in the input compared to not including. We expect that by training the model to embed articles that are close to each other in the hierarchical structure also close to each other in the embedding space, we can more effectively incorporate hierarchical structure and therefore improve performance.

5.1 Overview of Model

In our proposed model, we compute two loss functions which we call basic and contrastive, respectively. We employ as the baseline a model which is similar to Wehnert et al. (2021) [3] except we perform fine-tuning using the COLIEE training set by cosine similarity loss. We sum the cosine similarity loss and the contrastive loss, introduced by the proposed model, and the sum is our final objective function. The overview of the proposed method is shown in Figure 3. During training, the model converts the query and article into BERT representations and cal-

[2] We normalize the values of cosine similarity so that for each query, the pair of this query and the most similar article has similarity value of 1, and the pair of this query and the least similar article 0.

[3] Strictly speaking, our baseline also differs from Wehnert et al. (2021) as they append to the articles commentaries obtained by web crawling and the queries that are entailed by the articles, whereas we omit them in this study because the performance without fine-tuning is almost the same as what is reported in Wehnert et al. (2021) if we do not include them

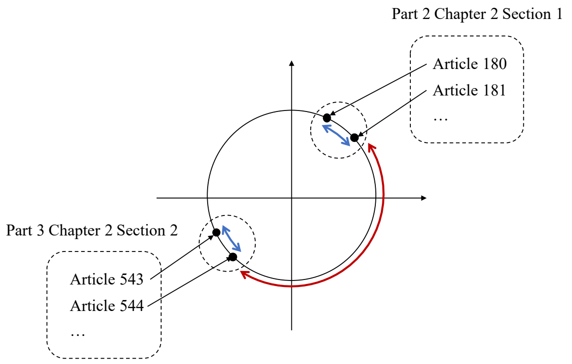


Figure 4: Idea of contrastive learning using hierarchical structure of law. The articles in the same section are pulled closer (blue arrows) while the those in different sections are pushed away (red arrow).

culates the cosine similarity loss from their cosine similarity. On the other hand, the contrastive loss is computed to bring the articles in the batch that belong to the same section closer together. The sum of the two losses is learned to be minimized. When making inference, cosine similarity scores between the query and the article are calculated using BERT and TF-IDF, which are each normalized. We sum them together, and then further normalize the scores to determine if they are relevant by comparing them to the threshold value.

5.2 Loss Function

We give the definitions of the loss functions below. First, for each query-article pair, $\text{pair}_k = (q_{k_q}, a_{k_a})$, we define the binary label indicating relevance of the pair l_k as

$$l_k = \begin{cases} 1 & \text{if } a_{k_a} \in \mathbb{A}_{k_q} \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

The query q_{k_q} and the article a_{k_a} are converted into features and denoted by $\mathbf{q_{k_q}} = \text{SBERT}(q_{k_q}) \in \mathbb{R}^d$ and $\mathbf{a_{k_a}} = \text{SBERT}(a_{k_a}) \in \mathbb{R}^d$, respectively, where $\text{SBERT}(x)$ is the feature corresponding to the CLS token when x is encoded by sentence-BERT, and d is the number of dimensions of the final layer of sentence-BERT.

5.2.1 Cosine Similarity Loss

Given N randomly sampled triplets of query representation, article representation, and the binary label indicating relevance of the two, cosine similarity loss is computed as follows:

$$\mathcal{L}^{cos} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_k^{cos}, \quad (2)$$

where

$$\mathcal{L}_k^{cos} = \begin{cases} 1 - \cos(\mathbf{q}_{k_q}, \mathbf{a}_{k_a}) & \text{if } l_k = 1 \\ \max(\cos(\mathbf{q}_{k_q}, \mathbf{a}_{k_a}), 0) & \text{otherwise} \end{cases} \quad (3)$$

Minimizing this loss function means penalizing positive cases for a lower similarity, and negative cases for a higher similarity.

5.2.2 Contrastive Loss

Figure 4 describes our idea for contrastive learning. Contrastive learning is performed with the pair of relevant article and an article in the same section as the relevant article as the positive example and the relevant article and an article randomly selected from all the relevant article as the negative example. The contrastive loss is calculated as follows. First, the set of representations of articles in the batch

$$\mathbb{A}_{\text{BATCH}}^{\text{EMB}} = \{\mathbf{a}_{b_1}, \mathbf{a}_{b_2}, \ldots, \mathbf{a}_{b_N}\}$$

is partitioned into groups

$$\mathbb{A}_{\text{BATCH}}^{\text{EMB}} = \bigcup_i \mathbb{S}_i$$

by section that they belong to. That is,

$$\mathbb{S}_i := \{\mathbf{a} \in \mathbb{A}_{\text{BATCH}}^{\text{EMB}} \mid \text{section}(\mathbf{a}) = i\},$$

where $\text{section}(\mathbf{a}) = i$ means article $\mathbf{a}$ belongs to section i.

Then, we construct triplets for contrastive learning. In summary, we (i) generate all possible pairs of articles that belong to the same section within the batch; then, (ii) to each pair, append a randomly selected negative example, which is an article from a different section, to form a triplet. Below is a more rigorous description. We let

$$\mathbb{C}_i := \{(\mathbf{a}_s, \mathbf{a}_t) \mid \mathbf{a}_s, \mathbf{a}_t \in \mathbb{S}_i; s < t\}$$

denote the set of all possible pairs of elements in $\mathbb{S}_i$. Also, let $\mathbb{C} := \bigcup_i \mathbb{C}_i$ and to each pair in $\mathbb{C}$, supply an article randomly chosen from a different section to form a triple, and denote the set of all these triple by $\tilde{\mathbb{C}}$. In other words

$$\tilde{\mathbb{C}} := \{(\mathbf{a}_s, \mathbf{a}_t, \mathbf{a}_{x(s,t)}) \mid (\mathbf{a}_s, \mathbf{a}_t) \in \mathbb{C}\}$$

where $x(s,t)$ is a random variable that takes a value in the set $\mathbb{I} := \{i \mid \mathbf{a}_i \in \mathbb{A}; \mathrm{section}(\mathbf{a}_i) \neq \mathrm{section}(\mathbf{a}_s) = \mathrm{section}(\mathbf{a}_t)\}$ according to the discrete uniform distribution. Hereafter, for simplicity, the notation of the elements of $\tilde{\mathbb{C}}$ will be changed as follows. Let $N_C := |\tilde{\mathbb{C}}|$ and for each $i \in \{1, 2 \ldots, N_C\}$, the i-th element of $\tilde{\mathbb{C}}$ is denoted by $(\mathbf{a}_i, \mathbf{a}_i^+, \mathbf{a}_i^-)$.

Then, the contrastive loss is

$$\mathcal{L}^{cont} = \\ -\log \frac{\mathrm{sim}(\mathbf{a}_i, \mathbf{a}_i^+)}{\sum_{j=1,\neq i}^{N_C} \mathrm{sim}(\mathbf{a}_i, \mathbf{a}_j^+) + \mathrm{sim}(\mathbf{a}_i, \mathbf{a}_j^-)} \quad (4)$$

where $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \exp(\cos(\mathbf{u}, \mathbf{v})/\tau)$ and τ is the temperature parameter. Intuitively, minimizing this loss function means making the inner product of $\mathbf{a}_i$ with the positive example larger than the inner product with the negative example and other vectors in the batch.

The final loss function of the model is the sum of cosine similarity loss and scaled contrastive loss $\mathcal{L} = \mathcal{L}^{cos} + \alpha\mathcal{L}^{cont}, \alpha \in (0, 1)$.

6 Experiments

As in the COLIEE competition, we employed the recall, precision, F2 values, and the percentage of correct answers in the top ten articles predicted by the model (R@10), as evaluation metrics. In light of the practical goal of improving access to legal information for non-specialists, as mentioned in the introduction, it is more important that the top predictions include possible legal topics to which a query may relate than to present irrelevant topics. Therefore, we pay particular attention to R@10. These evaluation metrics are calculated for each query and averaged over all queries.

6.1 Settings

Wehnert et al. (2021) uses a Sentence-BERT model pretrained with general paraphrase data [4] to make predictions, without fine-tuning it with COLIEE training data. In order to make a fair comparison, we compare our method to a baseline that computes the cosine similarity loss in the same manner as Wehnert et al. (2021) and train it on COLIEE data set before making a prediction.

Preliminary experiments using the validation data showed that the baseline setting tended to improve scores almost steadily up to about the 9th epoch, but not much thereafter. Therefore, we expected that introducing control learning from this point would be effective, and compared (1) a model with 18 epochs of learning in the baseline setting and (2) a model with 9 epochs of learning with the baseline setting and then 9 epochs of learning with the proposed setting.

The number of epochs is 18, batch size is 256, optimization algorithm is Adam (Kingma and Ba, 2015), and the learning rate is $1e - 6$, and hyperprameters regarding optimization other than the learning rate are set as recommended by (Kingma and Ba, 2015) for both the baseline and Ours. Learning rate is reset to the initial value every 3 epochs. As for parameters specific to contrastive learning in Ours, $\alpha = 1e - 7, \tau = 20$. We report the average of 3 trials along with their standard deviations, in the form of (average value) ± (standard deviation).

6.2 Results

The results of experiments are shown in Table 3. As for F2 and precision, we do not observe a significant difference in the performance of Ours and the baseline. However, Ours performs better than the baseline in terms of R@10 and recall. R@10 and recall are the main focus of our study since our purpose is to build a system that provides non-expert users relevant legal topics to the query. Note that we have compared our method to a baseline stronger than previous state-of-the-art, so that we can differentiate the effectiveness of contrastive learning from that of the plain fine-tuning. The results have shown a solid evidence that contrastive learning is effective.

6.3 Discussion

We test the hypothesis that when a relevant article with a high similarity to the query and an relevant article with a low similarity to it are in the same section, then bringing the representations of these arti-

[4] sentence-transformers/paraphrase-distilroberta-base-v1

Table 3: Comparison with the baseline. Wehnert et al. (2021) indicates the scores reported in the paper. (Wehnert et al., 2021)† is replication by us. The scores of the baseline and Ours is the average of three trials and standard deviation.

model	F2	Rec.	Pre.	R@10
(Wehnert et al., 2021)	73.02	77.78	**67.49**	81.20
(Wehnert et al., 2021)†	72.25	**82.09**	48.81	83.33
Baseline	74.87 ± 0.87	79.22 ± 1.20	61.47 ± 2.22	87.04 ± 0.00
Baseline + Contrastive (Ours)	**75.53** ± 0.79	80.66 ± 0.71	60.26 ± 1.82	**88.48** ± 0.71

Table 4: R@10 for each division by TF-IDF similarity

model	Hard R@10	Hard-Anchored R@10	Easy R@10
(Wehnert et al., 2021)†	43.18	45.00	**100.00**
Baseline	56.82 ± 0.00	70.00 ± 0.00	**100.00** ± 0.00
Baseline + Contrastive (Ours)	**60.61** ± 1.31	**78.33** ± 2.89	**100.00** ± 0.00

cles closer together is effective in improving performance.

In the test data, there are 101 pairs of a query and relevant article, of which 44 are hard and 57 are easy. 20 of the hard articles are in the same section as the article that has the highest cosine similarity with the query (called hard-anchored). R@10 for each division is shown in Table 4.

In Hard-Anchored, the advantage of Ours over the baseline is especially pronounced. This explains why contrastive learning is effective. That is, a relevant article with a low similarity with the query becomes more similar to the query by being brought closer to a relevant article with a high similarity with the query.

Figure 5 shows hard-anchored cases that could not be obtained with the baseline method in the top 10 prediction but could be with the proposed method. The article called ”Relevant Article” is the one which became obtainable and ”Anchor Article” is its same-section peer with high n-gram similarity to the query; the n-grams with underline highly overlap between the Query and Anchor Article.

On the other hand, the cases which could not be obtained either by the baseline or proposed method involved complicated coreference resolution and hypernym detection. Examples are shown in Figure 6. This implies that a limitation of our method is that it still has not overcome the difficulty of capturing the correspondence between general concepts and specific examples, which we shall consider in future work. We saw no case where the baseline method successfully obtained a relevant article in its top 10 prediction but the proposed method failed. This indicates that the proposed method has achieved improvement in some hard cases without any sacrifice in terms of R@10.

7 Conclusion

In this study, to address the difficulty with previous work to classifying relevant articles with low n-gram similarity to the query as relevant in a legal information retrieval task, we focused on the fact that many such relevant articles are hierarchically close to relevant articles with high n-gram similarity to the query, and proposed supervised contrast learning using hierarchical information. Experimental results show that the proposed method outperforms previous methods, especially in classifying articles with low n-gram similarity as correct answers.

Acknowledgements

We thank the organizers of the COLIEE for constructing this dataset.

References

Houda Alberts, Akin Ipek, Roderick Lucas, and Phillip Wozny. 2021. Coliee 2020: Legal information retrieval and entailment with legal embeddings and boosting. In Naoaki Okazaki, Katsutoshi Yada, Ken Satoh, and Koji Mineshima, editors, *New Frontiers in*

Artificial Intelligence, pages 211–225, Cham. Springer International Publishing.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*.

Carole D. Hafner. 1980. Representation of knowledge in a legal information retrieval system. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR '80, page 139–153, GBR. Butterworth amp; Co.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Prannay Khosla, Teterwak Piotr, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. In *NeurIPS*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *ACM Trans. Inf. Syst.*, 19(3):242–262, jul.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2020. Sparse, dense, and attentional representations for text retrieval.

Yixiao Ma, Yunqiu Shao, Bulou Liu, Min Zhang, and Shaoping Ma. 2021. Retrieving legal cases from a large-scale candidate corpus. In *ICAIL/COLIEE*.

Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2365, Austin, Texas, November. Association for Computational Linguistics.

Ha-Thanh Nguyen, Phuong Nguyen, Thi-Hai-Yen Vuong, Quan Bui, Chau Nguyen, Binh Dang, Vu Tran, Minh Nguyen, and Ken Satoh. 2021. Jnlp team: Deep learning approaches for legal processing tasks in coliee 2021. In *ICAIL/COLIEE*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert.

Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripa Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2021. Aila 2021: Shared task on artificial intelligence for legal assistance. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 12–15, New York, NY, USA. Association for Computing Machinery.

Juliano Rabelo, Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, and Ken Satoh. 2021. Summary of the competition on legal information extraction/entailment (coliee) 2021. In *ICAIL/COLIEE*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.

Guilherme Rosa, Ruan Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3501–3507. International Joint Conferences on Artificial Intelligence Organization, 7. Main track.

Hsuan-Lei Shao, Yi-Chia Chen, and Sieh-Chuen Huang. 2021. Bert-based ensemble model for statute law retrieval and legal information entailment. In Naoaki Okazaki, Katsutoshi Yada, Ken Satoh, and Koji Mineshima, editors, *New Frontiers in Artificial Intelligence*, pages 226–239.

Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1):65–87, mar.

Ellen M. Voorhees. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Sabine Wehnert, Viju Sudhi, Shipra Dureja, Libin Kutty, Saijal Shahania, and Ernesto W. De Luca. 2021. Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In *ICAIL/COLIEE*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. Cail2019-scm: A dataset of similar case matching in legal domain.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of bert for ad hoc document retrieval.

Query If the contract of sale stipulates that F, who was not born at the time of the conclusion of the contract, is to acquire the ownership of X, the contract of sale is invalid.

Relevant Article Article 537 (1) If one of the parties promises in a contract to render a certain performance to a third party, the third party has the right to claim that performance directly from the obligor. ... (Omitted)

Anchor Article Article 548-4 (1) In the following cases, a preparer of the standard terms of contract may, by amending the standard terms of contract, modify the terms of the contract without making separate agreements with each of the counterparties and deem that the parties have agreed to the amended provisions of the standard terms of contract: ... (Omitted)

Query If a third-party collateral provider paid a secured claim, the third-party collateral provider may exercise the secured claim acquired through subrogation without requirement for perfection.

Relevant Article Article 500 The provisions of Article 467 apply mutatis mutandis in the case referred to in the preceding Article (unless a person with a legitimate interest in making performance is subrogated to the claim of the obligee).

Anchor Article
Article 501 (1) A person that is subrogated ... (Omitted) ... (i) a third party acquirer (meaning a person that has acquired from the obligor the property that is the subject of security; hereinafter the same applies in this paragraph) is not subrogated to the claim of the obligee in relation to any guarantors or third-party collateral providers;... (Omitted)

Figure 5: Hard-Anchored cases which became obtainable by contrastive learning

Query A took the jewelry that B had forgotten, believing without negligence that it belonged to A. In this case, A may not obtain the ownership of the jewelry by good faith acquisition.

Relevant Article Article 192 A person that commences the possession of movables peacefully and openly by a transactional act acquires the rights that are exercised with respect to the movables immediately if the person possesses it in good faith and without negligence.

Query If D owes C a debt (Y) of 300000 yen that is set off against the debt (X), and D demands payment of 600000 yen from A while C does not use a set-off for the debt (Y), A may refuse to pay 200000 yen of that debt.

Relevant Article Article 439 (1) If one of the joint and several obligors has a claim against the obligee and invokes a set-off, the claim is extinguished for the benefit of all joint and several obligors. (2) Until the joint and several obligor that has the claim referred to in the preceding paragraph invokes a set-off, other joint and several obligors may refuse to perform the obligation to the obligee only to the extent of that joint and several obligor's share of the obligation.

Figure 6: Hard case which is not obtainable either by baseline or proposed method

Approaching Neural Chinese Word Segmentation as a Low-Resource Machine Translation Task

Pinzhen Chen **Kenneth Heafield**
School of Informatics, University of Edinburgh
{pinzhen.chen, kenneth.heafield}@ed.ac.uk

Abstract

Chinese word segmentation has entered the deep learning era which greatly reduces the hassle of feature engineering. Recently, some researchers attempted to treat it as character-level translation, which further simplified model designing, but there is a performance gap between the translation-based approach and other methods. This motivates our work, in which we apply the best practices from low-resource neural machine translation to supervised Chinese segmentation. We examine a series of techniques including regularization, data augmentation, objective weighting, transfer learning, and ensembling. Compared to previous works, our low-resource translation-based method maintains the effortless model design, yet achieves the same result as state of the art in the constrained evaluation without using additional data.

1 Introduction

Chinese text is written in characters as the smallest unit, and it has no explicit word boundary. Therefore, Chinese word segmentation (CWS) serves as upstream tokenization and disambiguation for Chinese language processing. The task is often viewed as sequence labelling, where each character receives a label indicating its relative position in a segmented sequence (e.g. whether the character is at the word boundary). While traditional machine learning methods have attained strong results, recent investigations focus on neural networks given their rise in the entire NLP field. Distinctive to sequence labelling, Shi et al. (2017) first treat CWS as neural machine translation (NMT). Nonetheless, Zhao et al. (2018) point out that without extra resources, all previous neural methods are not yet comparable with the non-neural state of the art from Zhao and Kit (2008); the NMT practice is even behind.

We note two advantages of treating the task as neural translation: 1) the entire input sentence is encoded before making any segmentation decision; 2) such a model jointly trains character embeddings with sequence modelling. Thus, we try to bridge the gap between the translation-based approach and state-of-the-art models, using low-resource techniques commonly seen in NMT. The translation-based method can be easy to adopt without the need for feature extraction and model modifictaion. Although NMT is known to be data hungry, our approach is able to achieve competitive results in the constrained evaluation scenario, where introducing extra data is forbidden. In specific, when benchmarked on the second CWS bakeoff (Emerson, 2005), our system reaches the top of the MSR leaderboard and achieves a strong result on the PKU dataset.

2 Related Work

Chinese segmentation is traditionally tackled as sequence labelling, which predicts whether each input character should be split from neighbouring characters (Xue, 2003). Earlier approaches relied on conditional random fields or maximum entropy Markov models (Peng et al., 2004; Ng and Low, 2004). Zhao and Kit (2008) leveraged unsupervised features to attain state-of-the-art results in the data-constrained track.

Recent research has shifted towards neural networks: feed-forward, recurrent and convolutional (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a,b; Wang and Xu, 2017). Without external data, these models did not surpass the best non-neural method, but instead, provided great ease of data engineering. Researchers also studied better representations for segments and characters, as well as the incorporation of external resources (Liu et al., 2016; Zhou et al., 2017; Yang et al., 2017). By carefully tuning model configurations, Ma et al. (2018) achieved strong results. The task can also be done through learning to score global word segmentation schemes given characters (Zhang and Clark, 2007, 2011; Cai and Zhao, 2016; Cai et al., 2017). On top of this, Wang et al. (2019) proved that it is beneficial to integrate unsupervised segmentation. A recent work used a modified Transformer for sequence tagging to attain the same results as the state of the art (Duan and Zhao, 2020).

The most relevant to our research is Shi et al. (2017)'s suggestion to formalize Chinese word segmentation as character-level neural machine translation. It differs from global segmentation scoring in that the NMT directly generates Chinese characters with delimiters. It can also be equipped with post-editing that adds back characters omitted by the model. Later, Wei et al. (2019) restrict the NMT decoding to follow all and only the input characters. This proposal, together with existing NMT toolkits, eases the model design and implementation for neural Chinese segmentation. However, even with external resources, the two systems are inferior to the previous works concerning performance. This encourages us to explore low-resource techniques to enhance the NMT-based approach.

3 Methodology

An NMT model is trained to minimize the sum of an objective function L over each target sentence $y_0^n = y_0, y_1, ..., y_n$ given a source sentence X. We list below per-character conditional cross-entropy as an example:

$$L = -\frac{1}{n}\sum_{i=1}^{n}\log P(y_i|y_0^{i-1}, X) \qquad (1)$$

Following Shi et al. (2017), we make use of character-level NMT, and add an extra delimiter token "$\langle D\rangle$" to the target vocabulary. The delimiter token on the target side implies that the previous and next words are separated. To visualize, given an unsegmented sentence in characters "我会游泳", the model should output character-by-character "我$\langle D\rangle$ 会$\langle D\rangle$ 游泳" (English: I can swim). This in reality resembles how a human would read an unsegmented Chinese sentence.

We argue that NMT can model word segmentation well because the decoder has access to the global information in both decoder and attention states. Moreover, the output segmented characters may display stronger probabilistic patterns than the position labels do, resulting in more explicit modelling of the word boundary "$\langle D\rangle$". This characteristic is also robust to out-of-vocabulary words because NMT can freely "insert" the boundary delimiter anywhere to form words. Finally, this method does not require any alteration to the model architecture.

However, it poses a challenge when CWS is approached as NMT, that NMT usually performs poorly under a low-resource condition (Koehn and Knowles, 2017), which is exactly the case of Chinese segmentation datasets. A CWS corpus provides fewer than 100k sentences, whereas a typical translation task provides data at least one order of magnitude larger. To address this issue, we apply low-resource NMT practices: regularization and data augmentation. Then, we examine several other broadly used techniques.

3.1 Hyperparameter tuning

Hyperparameter tuning is often the first step to build a machine learning model. In the field of neural translation, Sennrich and Zhang (2019) show that carefully tuning hyperparameters results in substantial improvement in low-resource scenarios. In our case, we concentrate on regularization: label smoothing, network dropout, and source token dropout (Szegedy et al., 2016; Srivastava et al., 2014; Sennrich et al., 2016a). Additionally, we switch between GRU and LSTM, and increase the model depth (Hochreiter and Schmidhuber, 1997; Cho et al., 2014).

3.2 Objective weighting

The generic NMT objective function considers the loss from each target sentence or token equally. By adjusting the objective function we can make it weigh some components more than others, in order to better learn the desired part of the training data. It can be applied at the token or sentence level, for various purposes including domain adaptation and grammatical error correction (Chen et al., 2017; Wang et al., 2017; Yan et al., 2018; Junczys-Dowmunt et al., 2018b).

We propose to put more emphasis on the delimiter token in target sentences because they correspond to word boundaries directly. We weight delimiters k times as many as other tokens in the objective function, where k can be empirically determined on a validation set. The new token-weighted objective function L_{token} is as Equation 2, where the weight coefficient $\boldsymbol{\lambda}_i = k$ if y_i is a delimiter and $\boldsymbol{\lambda}_i = 1$ otherwise.

$$L_{token} = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda_i} \log P(y_i|y_0^{i-1}, X) \qquad (2)$$

3.3 Data augmentation

Data augmentation is widely adopted in NMT. The paradigm is to generate source side data from existing (monolingual) target side data (Sennrich et al., 2016b; Grundkiewicz et al., 2019), but this does not apply to CWS since there is no extra gold segmented data. Hence, we experiment with two methods that could suit CWS better: sentence splitting and unsupervised segmentation.

3.3.1 Sentence splitting

The surface texts of inputs and outputs are consistent in the NMT approach to CWS, with the only exception being the added delimiters. With a potential quality degrade, we assume that segmentation can be inferred locally, i.e. within a phrase instead of the whole sentence. It enables us to split a sentence into multiple shorter segments, with the gold segmentation unchanged. This can hugely expand the amount of training data, and reduce the input and output sequence lengths. In practice, we break full sentences down at comma and period symbols, since they are always separated from other characters.

3.3.2 Unsupervised segmentation

Both Zhao and Kit (2008)'s, and Wang et al. (2019)'s papers show that unsupervised segmentation helps supervised CWS. We use an external tool to segment our training data in an unsupervised way, to create augmented data (detailed later in Section 4.3). The data is utilized in two scenarios different from previous works: sentence-level weighting and transfer learning. The methods are depicted below.

Sentence weighting Weighting objective function at the sentence level can distinguish high- and low-quality training data. We designate our unsupervised segmentation result as low-quality augmented data, and the original training sentences as high-quality data. After combining them into a single training set, the high-quality data is weighted k times as much as the low-quality data. Equation 3 shows that sentence-weighted objective function $L_{sentence}$, where weight $\boldsymbol{\lambda} = k$ for gold sentences and 1 for augmented sentences. In contrast with Equation 2, the sentence weight is not token-dependant.

$$L_{sentence} = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda} \log P(y_i|y_0^{i-1}, X) \qquad (3)$$

Transfer learning It means to pre-train a model on high-resource data and then optimize it for a low-resource task. It often yields enhanced results over directly training on a small dataset (Zoph et al., 2016), as the knowledge learned from the high-resource task can be beneficial. Moreover, Aji et al. (2020) claim that starting a model from trained parameters is better than random initialization. We first train a model on the augmented data from an unsupervised segmenter, then further optimize it on the genuine training data.

3.4 Ensembling

An ensemble of diverse and independently trained and models enhances prediction. In our work, we combine models trained with different techniques and random seeds, and integrate a neural generative language model (LM) trained on the gold segmented training data. It works as follows: at each inference time step, all models' predictions are simply averaged to form the ensemble's prediction over the target vocabulary.

4 Experiments and Results

4.1 Task description

Evaluation takes place on the Microsoft Research (MSR) and Peking University (PKU) corpora in the second CWS bakeoff (Emerson, 2005).[1] The datasets are of sizes 87k and 19k, which are considered low-resource in machine translation tasks. Regarding preprocessing, our own training and validation sets are created randomly at a ratio of 99:1, from the supplied training data. We normalize characters, and convert continuous digits and Latin alphabets to "$\langle N\rangle$" and "$\langle L\rangle$" symbols without affecting segmentation.

There are both closed (constrained) and open tests in the CWS bakeoff. The former requires a system to only use the supplied data. Since we aim to strengthen the translation-based approach itself, we select the closed test condition and compare with other papers that report closed test results. The evaluation metric F1 (%) is calculated by the script from the bakeoff. We test different

[1] sighan.cs.uchicago.edu/bakeoff2005.

techniques on MSR and apply the best configurations to PKU without further tuning.

	$\text{drop}_{\text{state}}$	best loss	
drop_{src} = 0	0	0.0333	
	0.1	0.0271	
	0.2	**0.0262**	✓
	0.3	0.0272	
	0.4	0.0303	

	drop_{src}	best loss	
$\text{drop}_{\text{state}}$ = 0.2	**0**	**0.0262**	✓
	0.15	0.2081	
	0.3	0.4496	

(a) Experiments on two dropout methods. drop_{src} indicates entire source word dropout and $\text{drop}_{\text{state}}$ indicates dropout between RNN states.

	label smoothing	best loss	
drop_{src}= 0, $\text{drop}_{\text{cell}}$= 0.2	**0**	**0.0262**	✓
	0.1	0.1161	
	0.2	0.2220	

(b) Experiments on label smoothing.

cell	encoder depth	decoder depth	best loss	
GRU	1	1	0.0262	✓
	1	2	0.0251	
	2	1	0.0261	
	2	2	0.0264	
	3	3	0.0276	
	4	4	0.0268	
LSTM	1	1	0.0286	

(c) Experiments on model depth and the RNN type. No obvious winner is observed.

Table 1: Hyperparameter searches.

4.2 Baseline with regularization

We start with a 1-layer bi-directional GRU with attention (Bahdanau et al., 2015) containing 36M parameters. Adam (Kingma and Ba, 2015) is used to optimize for per-character (token) cross-entropy until the cost on the validation set stalls for 10 consecutive times. We set the learning rate to 10^{-4}, beam size to 6, and enable layer normalization (Ba et al., 2016). Since the model input and output share the same set of characters, we use a shared vocabulary and tied embeddings for source, target, and output layers (Press and Wolf, 2017). Training such a model on the MSR dataset takes 5 hours on a single GeForce GTX TITAN X GPU with the Marian toolkit (Junczys-Dowmunt et al., 2018a).[2]

Regarding hyperparameter selection, we always select the best settings based on the loss on the validation set. The tuning procedures are reported in Table 1. We see that a small dropout of 0.2 is helpful; source token dropout and label smoothing both cause adverse effects. Changing model depth and switching from GRU to LSTM make a negligible impact, so we stick to the single-layer GRU.

The first row in Table 4 shows that our carefully-tuned baseline achieves an F1 of 96.8% on the MSR test set. Next, we find that weighting delimiters twice as other tokens brings a 0.1% improvement. Delimiter weight tuning is presented in Table 2. These numbers already outperform previous translation-based works.

weight λ on delimiters	best loss	
1 (no weighting)	0.0262	
1.5	0.0197	
2	**0.0191**	✓
4	0.0204	
10	0.0210	
50	0.0253	

Table 2: Experiments on delimiter (word) weighting. λ is the weight on the delimiter, and other words are always given a weight of 1.

weight λ on original data	best loss	
1 (no weighting)	0.0462	
2	0.0346	
5	0.0309	
10	0.0268	
20	0.0227	
40	**0.0226**	✓
100	0.0230	
200	0.0245	
only genuine data	0.0268	

Table 3: Experiments on weighting augmented and original data. λ represents the weight on original sentences; augmented data always have a weight of 1.

4.3 Leveraging augmented data

Sentence splitting is done on both sides of the training and validation sets. Test sentences are

[2] https://github.com/marian-nmt/marian.

Techniques	F1 (%)
baseline w/ regularization (base)	96.8
base + delimiter weight	96.9
base + sentence splitting (split)	97.1
base + split + unsupervised + transfer	97.1
base + split + unsupervised + weight	**97.3**
2 × baseline	97.2
2 × transfer + 2 × weight + LM	**97.6**

Table 4: F1 of our techniques on MSR test set.

split, segmented by the model, and then concatenated, ensuring a consistent evaluation outcome. This leads to a better F1 of 97.1%, thanks to a 3-fold increase in data size to 257k for MSR.

We employ the segmental language model (Sun and Deng, 2018) for unsupervised segmentation.[3] We used the MSR model optimized on the training, validation, and test sets with a maximum word length of 4. Since the system is fully unsupervised, it is fair to include the test set; yet we only apply it to our training split to generate augmented data. In this way, no external resource is introduced. While transfer learning brings no gain, sentence-level weighting lifts the overall score to 97.3%, as shown in Table 4. We see that the cost on the validation set improves, and then degrades as sentence weight gets larger; the best sentence weight is determined to be 40 for MSR. The detailed weight selection process is described in Table 3.

4.4 Ensembling

During decoding, all models' predictions are averaged to produce an output token at each step. We first test an ensemble consisting of two baselines. Next, we combine two transfer learning models, two sentence-weighting models, and a character RNN LM. The LM has the same architecture as our NMT decoder. It is optimized for perplexity on the segmented side of the train set. Ensembling is done in one shot without tuning weights and it achieves the highest F1 of 97.6%.

5 Results and Analysis

In addition to MSR test, we keep the best hyperparameters determined on the MSR corpus unchanged, and run the same set of experiments on the PKU dataset.

System		MSR	PKU
non-neural	Zhao and Kit, 2008	**97.6**	95.4
	Zhang and Clark, 2011	97.3	94.4
neural	Pei et al., 2014	94.4	93.5
	Cai and Zhao, 2016	96.4	95.2
	Wang and Xu, 2017	96.7	94.7
	Cai et al., 2017	97.0	95.4
	Zhou et al., 2017	97.2	95.0
	Ma et al., 2018	**97.5**	95.4
	Wang et al., 2019	97.4	**95.7**
	Duan and Zhao, 2020	**97.6**	95.5
NMT-based	Shi et al., 2017	94.1	87.8
	+ external resources†	96.2	95.0
	Wei et al., 2019†	94.4	92.0
	Our best single model	97.3	95.0
	Our best ensemble‡	**97.6**	95.4

† The results are advantaged as extra resources are used.
‡ 97.61±0.16 on MSR and 95.43±0.38 on PKU, with $p < 0.05$ using bootstrapping (Ma et al., 2018), detailed in Appendix A.

Table 5: Previous and our systems' F1 (%) on MSR and PKU corpora under the constrained condition.

Table 5 compares our MSR and PKU results with previous papers. Our best single models are remarkably ahead of other NMT-based methods. With ensembling, our result on MSR ties with state of the art, showing that empirically neural methods can reach the top without external data. However, as data size significantly drops in the case of PKU, we observe a declined performance and larger variance on the PKU dataset. This is expected as NMT is known to be sensitive to a smaller data size.

Regarding regularization, we discover that low-resource NMT techniques are not always constructive for CWS. Dropping out source tokens is harmful because CWS is not a language generation task and the decoder output heavily relies on the input. A similar rationale explains why label smoothing causes rocketing cross-entropy: unlike language generation where a variety of outputs are accepted, for CWS there is always just one single correct scheme. Smoothing out the decoder probability distribution results in confusion.

Further, unsupervised data augmentation with weighting achieves the best single-model result. We suggest a possible reason: the augmented data has the same source side as the original data, but a noisier target side. When weighted appropriately, the noise might act as a smoothing tech-

[3] Their code and released models: github.com/edward-sun/slm.

nique for sequence modelling, especially in the low-resource condition (Xie et al., 2017). From the transfer learning aspect, pre-training on the augmented data does not lead to a higher number than starting from a randomly initialized state.

6 Conclusion

Our low-resource translation-based approach to Chinese word segmentation achieves strong performance and is easy to adopt. Data augmentation, objective weighting and ensembling are the most favourable. In future, it is worth extending this perspective to word segmentation of other languages, as well as re-basing it on Transformer models.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9117. The work reflects the view of the authors and not necessarily that of the funders.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *ACL*, pages 7701–7710.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for Chinese. In *ACL*, pages 409–420.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. In *ACL*, pages 608–615.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. 2015a. Gated recursive neural network for Chinese word segmentation. In *ACL-IJCNLP*, pages 1744–1753.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015b. Long short-term memory neural networks for Chinese word segmentation. In *EMNLP*, pages 1197–1206.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Sufeng Duan and Hai Zhao. 2020. Attention is all you need for Chinese word segmentation. In *EMNLP*, pages 3862–3872.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *ACL*, pages 116–121.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018b. Approaching neural grammatical error correction as a low-resource machine translation task. In *NAACL*, pages 595–606.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. Exploring segment representations for neural segmentation models. In *IJCAI*, pages 2880–2886.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art Chinese word segmentation with bi-LSTMs. In *EMNLP*, pages 4902–4908.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *EMNLP*, pages 277–284.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *ACL*, pages 293–303.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING*, pages 562–568.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *EACL*, pages 157–163.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *WMT*, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *ACL*, pages 211–221.

Xuewen Shi, Heyan Huang, Ping Jian, Yuhang Guo, Xiaochi Wei, and Yi-Kun Tang. 2017. Neural chinese word segmentation as sequence to sequence translation. In *Chinese National Conference on Social Media Processing*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(56):1929–1958.

Zhiqing Sun and Zhi-Hong Deng. 2018. Unsupervised neural word segmentation for Chinese via segmental language modeling. In *EMNLP*, pages 4915–4920.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for Chinese word segmentation. In *IJCNLP*, pages 163–172.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *EMNLP*, pages 1482–1488.

Xiaobin Wang, Deng Cai, Linlin Li, Guangwei Xu, Hai Zhao, and Luo Si. 2019. Unsupervised learning helps supervised neural word segmentation. In *AAAI*, pages 7200–7207.

Yuekun Wei, Binbin Qu, Nan Hu, and Liu Han. 2019. An improved method of applying a machine translation model to a chinese word segmentation task. In *ICANN*, pages 44–54.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. In *ICLR*.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics and Chinese Language Processing*.

Shen Yan, Leonard Dahlmann, Pavel Petrushkov, Sanjika Hewavitharana, and Shahram Khadivi. 2018. Word-based domain adaptation for neural machine translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 31–38.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *ACL*, pages 839–849.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In *ACL*, pages 840–847.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Hai Zhao, Deng Cai, Changning Huang, and Chunyu Kit. 2018. Chinese word segmentation: Another decade review (2007-2017). In *Frontiers of Empirical and Corpus Linguistics*, pages 139–162.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. In *EMNLP*, pages 647–657.

Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2017. Word-context character embeddings for Chinese word segmentation. In *EMNLP*, pages 760–766.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *EMNLP*, pages 1568–1575.

A Results with a confidence interval

We report our final score with a confidence interval since the top results are very close. As there is only one test set, we create another 599 test sets of the same size as the original one, through resampling with replacement. Our best system obtains an F1 of 97.61±0.16 on the MSR dataset and 95.43±0.38 on the PKU dataset with 95% confidence (2 standard deviations).

Attitudes towards pedagogical code-switching: A verbal guise approach

Andrea Althea G. Aballa[1], Chyme Fresca M. Colinares[1], Danielle Majella C. Milanes[1], John Chael D. Wang[1]
Edward Jay M. Quinto[2], John Christopher D. Castillo[2]
[1]Mapua University, Manila, Philippines
[2]School of Social Sciences and Education, Mapua University, Manila, Philippines
aagaballa@mymail.mapua.edu.ph,
cfcolinares@mymail.mapua.edu.ph,dmcmilanes@mymail.mapua.edu.ph,
jcdwang@mymail.mapua.edu.ph, ejmquinto@mapua.edu.ph,
jcdcastillo@mapua.edu.ph

Abstract

The role of pedagogical code-switching (henceforth CS) has been an arising topic of inquiry across the globe; hence, the global surge in bilingualism significantly opened an opportunity to use CS as a resource in class sessions for effective learning. This study aims to distinguish the factors that influence the attitudes of the participants toward Tagalog and English pedagogical CS and identify the significant differences between English, Tagalog, and CS among Filipinos. Anchored on Myers-Scotton's (1993) Markedness model framework, this quasi-experimental study aims to identify the attitudes towards pedagogical CS compared to monolingual English and monolingual Tagalog. To do this, the researchers used the verbal Guise technique (VGT), an innovative approach used to study attitudes, which is composed of three speakers for each language (English, Tagalog, and CS), and integrated it within a google forms questionnaire that has a 4-point Likert Scale adapted from the study of Valerio (2015), which were then given and listened to the 784 purposively sampled senior high school and college students within the different universities in the Philippines. The researchers then analyzed the data using a non-parametric statistical treatment known as Friedman's ANOVA and Kendall coefficient of concordance, which compares three groups without the independent-dependent variable relationships.

1. Introduction

The role of pedagogical code-switching (henceforth CS) has been an arising topic of inquiry across the globe. The global surge in bilingualism significantly opened an opportunity to use CS as a resource in class sessions for effective learning (Soundiraraj, 2013). According to Feng and Chen (2009), direct relationships between emotional aspects are visualized when language is learned and comprehended through time. As a branch of linguistics, this indicates that CS may be studied through an individual's attitude, preferences, and cognitive values. This study then branched different variables, starting with the positive attitudes toward language learning, which Bhaskar and Soundiraraj (2013) concluded that could boost students' study, but negative attitudes can eventually block it, as well as different reasons concerning why teachers and students utilize CS while associating with one another, ie. reasons behind it were arranged into four regions: starting a new topic, understanding, emphasis, and lack of vocabulary (Gulzar, 2010), and the use of verbal guise technique (VGT) to gather data from a participants' attitude and behavioral approach towards the field of linguistics and any other language-related studies.

According to Goulet (1971), Tagalog-English (Taglish) is widely used throughout the Philippines' Tagalog-speaking region and is often regarded as "the typical acceptable conversation style of speaking and writing." Despite the extensive use and favorable attitudes toward this form of communication, Tagalog-English code-switching is underrepresented in the CS literature (Labitigan, 2013). Today, Tagalog-English CS can be mostly seen among the people of the Philippines who are educated, middle- and upper-class urbanites. However, Filipinos are likewise apt to quit one language for the social benefits of another when it comes to linguistic attitudes. At the time of Lesada (2017) in the Philippines, it was reported that when a family was visited, they intended to raise their young children entirely in English, excluding their native Waray, Cebuano, and Filipino from their home. When questioned why they decided, they stated that it would be better for their children and family in terms of education, socialization, and economics. This prevalent mindset demonstrates that English is still the language of prestige, and it exemplifies the concept of linguistic fluidity once more. Following the proclamation of the bilingual policy, disputes erupted about whether Taglish CS should be accepted as a medium of communication in academic institutions, specifically for imparting classroom instructions and academic discourse in general. In addition, it was said in the study of Bautista (2004) that Taglish does not inspire enthusiasm among instructors. In the study of Labor (2016), it was mentioned that CS is synonymous with incompetence and interference (Gumperz, 1982). Furthermore, Taglish was a "corruption of English and Tagalog languages," and its use as a communication medium revealed the speaker's lack of linguistic understanding, even though most of them were middle-class and educated (Bautista, 2004; and Flores, 2019). Nevertheless, Bautista believes that CS improves communicative efficiency because it is the quickest, easiest, and most effective communication method. As a result, the Philippines offers a veritable natural laboratory for assessing the communicative, pedagogical, and sociolinguistic benefits and drawbacks of CS (Bautista, 1991). In the Philippines, the DepEd's Mother Tongue Based - Multilingual Education (MTB-MLE) policy objectives and projected outcomes are comprehensive and research-based, but execution is a challenge. Based on the study of Burton (2013), a few countries have sought to implement multilingual education programs at the national level, but have faltered due to the vastness of the challenge, for instance, Bolivia, Ethiopia, Peru, South Africa. Top-down approaches, according to scholars, ignore the contextualized nature of language in communities (Kaplan, 1990; Martin-Jones and Saxena, 1995; Ricento and Hornberger, 1996). As a result, the viability of implementing MTB-MLE as a national strategy in the Philippines is being questioned. It is uncertain whether the research's scholarly findings will be applied to top-down policy circumstances. Previous research has shown the difficulty that mother tongue education initiatives face when they come up against local views that prefer English (Iyamu and Ogiegbaen, 2007). This is in direct opposition to the goals of MTB-MLE, and it raises the possibility of a dispute over its implementation. However, there are few studies done that focus on the attitudes towards Tagalog-English CS (Bautista, 2004; Burton, 2013; Labitigan, 2013; Odejar, Koutsoftas, and Marzan, 2016; Labor, 2016; Lesada, 2017; Flores, 2019), specifically in terms of utilizing the verbal guise technique (VGT). For instance, the studies of Bautista (1980 [1974], summarized in 1975), Marfil and Pasigna (1970), Palines (1981), Pimentel (1972), Sadicon (1978), and Sobolewski (1980, summarized in 1982) mainly attempted to define the structure of CS in the linguistic context in print and media corpora. Previous studies have identified CS outside of the pedagogical context as being directly opposed to pedagogical CS.

To address the aforementioned gaps, the present study aims to distinguish the factors that influence the participants' attitude toward Tagalog-English pedagogical CS and identify the attitudes and its significant differences between English, Tagalog, and CS among Filipinos. The researchers will utilize a unique and uncommonly used but credible indirect method of collecting data: verbal guise technique (VGT). This presents opportunities for researchers to extend knowledge from past studies that had repeatedly and dominantly used only direct means of approach such as surveys, questionnaires, and interviews. The study's outcome hopes to alleviate misconceptions towards CS while contributing to the discussion about the critical policies within MTB-MLE and the Filipino language. The practical value of the research stems from interactions that occur every day through diversified language variations. These languages are related to attitudes that express the cognitive abilities of an individual. Moreover, the behavioral aspect of attitude (BAA) is concerned with how people act and react in specific situations. With a deeper understanding of the attitudinal aspects in language and the emerging CS concept, sectors of education, business and industry, and mass media can benefit from simplifying a complex and malleable topic that is language.

With this, the researchers aim to answer the following research questions:

- What are the attitudes of the participants towards Tagalog, English, and Tagalog-English pedagogical CS?

- Is there a significant difference in the attitudes of the participants towards Tagalog, English, and CS?

2. Review of Related Literature

Language attitudes research has lately regained popularity among language academics as a result of shifting language policies in different countries. It is critical in the Philippines especially with the language policy being versatile. The Department of Education (DepEd) released an order that instructs institutions to use MTB-MLE wherein the use of the learners' first language was required to be utilized as the medium of instruction for all subject areas, except for Filipino and English being taught in different subjects. Previous research has demonstrated the challenges that mother tongue education programs face when confronted with local preferences for English (Iyamu and Ogiegbaen, 2007). This is in direct conflict with MTB-MLE's goals and it increases the risk of a controversy against its implementation. However, little research has been conducted on attitudes regarding Tagalog-English CS (Bautista, 2004; Burton, 2013; Labitigan, 2013; Odejar, Koutsoftas, and Marzan, 2016; Labor, 2016; Lesada, 2017; Flores, 2019), more so with the use of verbal guise technique (VGT). Language is a resource of building identity (Couplans 2007, Edwards 1999; Ladegaard, 2000; Meyerhoff, 2006) connected to motivation and attitude affecting CS. Though there are studies concentrating on CS, most of them are focused and developed with the use of other languages instead of Tagalog-English. This study is an attempt to fill the research gap of inadequate studies about Tagalog-English CS, particularly pedagogical CS. The majority of language-related studies that were referenced by the researchers are only focused on utilizing the methods of direct measures such as surveys, interviews, and scales. There are only a few research papers which use indirect methods such as MGT and VGT. VGT as a method focuses on behavioral manifestations of attitudes towards languages.

3. Framework

This research is based on Myers-Scotton's Markedness model (1993), which provides a useful framework for analyzing various types of CS as well as the fact that it serves a special purpose in multicultural and multilingual classrooms and is aligned with the researchers' data collection tool as VGT focuses more on the behavioral manifestations of attitudes. This framework is in favor of this study because bilingual speakers are aware of the social consequences of choosing a specific language in a given scenario, which is evaluated in terms of the marked versus unmarked opposition, as well as the degree to which it matches community expectations for the interaction type. Therefore, code selection is intentional in the Markedness Model since they are frequently made to fulfill specific communication goals.

4. Methodology

This section tackles the research design, setting and sampling technique, data collection, data gathering procedure, and data analysis used in the study.

Research Design

This study followed a quasi-experimental research design based on Campbell and Stanley's (1963) quasi-experiments because it establishes a causal association between an independent and dependent variable while assigning individuals to groups using non-random criteria (Cook and Campbell, 1979). The research design had an internal validity often higher than correlational studies but lower in actual experiments since the research design includes manipulating the independent variable without randomly assigning individuals to conditions or sequences of conditions, thus, quasi-experimental research eliminates the directionality problem (Price, 2015). This research design was previously preferred in studies, particularly those that looked at behaviors and attitudes (Adeeb, 1994; Sun and Eun, 2005; Rivera, 2015; Barnes, 2019). Utilizing the quasi-experimental research design, the researchers aimed to determine the attitudes and its significant differences between Tagalog, English, and CS in a pedagogical context among Filipino senior high school and college students.

Setting and Sampling Technique

The study was conducted in a tropical country in the Pacific Ocean that consists of 7,107 islands. Its institutions use MTB-MLE wherein the use of the learners' first language was required to be utilized as the medium of instruction for all subject areas, except for Filipino and English being taught in different subjects wherein several studies are done that challenges this (Iyamu and Ogiegbaen, 2007). As a country rich in languages, more than 100 distinct dialects, it is known to continue experiencing a period of language convergence, and borrowing from large languages such as English, Tagalog, as well as regionally important languages. As a melting pot of different cultures and languages, Filipinos are part of a veritable natural laboratory for assessing the communicative, pedagogical, and sociolinguistic benefits and drawbacks of CS (Bautista, 1991). The population of the study are Tagalog-English speakers in the Philippines. Among all Filipino teachers and students, the researchers selected participants who are currently in the senior high school and college level. Because motivation and attitude are so closely linked, research has discovered that they play a substantial influence in language learning. Their attitudes toward language acquisition can have a significant impact on the expected outcomes of classroom participation (Gardner, 1985; Krashen, 1982). The intended sample size is 1000 who are Tagalog-English senior high school and college students speakers. The participants will be sampled using purposive sampling. Purposive sampling (also known as judgment, selective or subjective sampling) is a sampling technique in which a researcher relies on his or her own judgment when choosing members of the population to participate in the study (Business Research Methodology, n.d.). This technique was deemed most appropriate because the researchers based the participants on certain criteria. This includes their educational level (senior high school or college), the age that should range from 16-24 years old, and should know how to speak Tagalog and English languages. First, analyze and prepare the set of criteria needed for the set of participants.

Data Collection

The instruments used in the study are in line with the methods of verbal guise technique. This includes a verbal-guise test and a questionnaire. Both instruments will be executed and recorded in English and Filipino language. The verbal guise test will be embedded within the questionnaire through google forms. Its purpose is to be able to gather data from one source of instrument. According to Lam (2006), questionnaires can be means of measuring and observing attitudes and perceptions of individuals. A questionnaire will be developed through a format of a 4-point Likert scale. It will be used by the individual respondent to evaluate the given items. The questionnaire contains four parts such as (1) the demographic profile of the participants; (2) the questions regarding their attitudes towards English and Filipino languages, and toward CS; (3) reasons for CS; (4) their perception on mother-tongue based multilingual education. For the first part of the questionnaire, the demographic profile of the participants will be described by their age, gender, study level, ethnic affiliation, and socioeconomic status. The second part of the questionnaire focuses on the attitudes of the participants toward the English language, attitudes of the participants toward the Filipino language, and attitudes of the participants toward CS. The third part deals with the reasons of the participants for CS. The last part of the questionnaire will be about the perception of the participants on mother-tongue-based multilingual education. Adapted from the study of Valerio (2015), the participants will be asked to rate each concern using a 4-point Likert Scale: (4) strongly agree, (3) agree, (2) disagree, (1) strongly disagree. The data was gathered using Google Forms, which included embedded audio of three professors discussing using the language of Tagalog, English, and CS. The speakers were chosen based on the same criteria: a female, 20-30 years of age and a senior high school or college professor with a master's degree in education. The usage of a microphone with modified audio levels of -12dB to -6dB was employed to produce a standardized quality and clear audio to minimize extraneous elements within recordings altering a participants' rating of the speakers. The poll was widely distributed by posting it on the researchers' Facebook account and sending emails to several universities in the Philippines, with agreement sought under the Data Privacy Notice and Ethical Considerations.

The following steps will carry out in this quasi-experimental research on the attitudes towards English, Tagalog, and Tagalog-English CS:

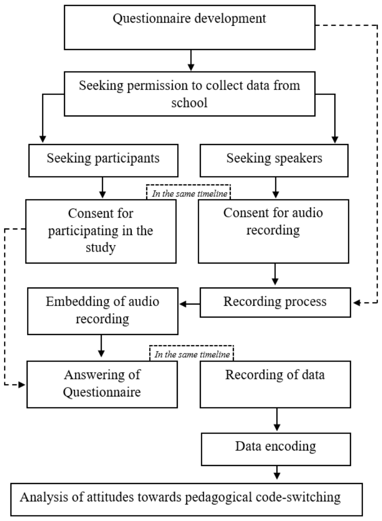

Questionnaire development
Seeking permission to collect data from school
Seeking participants
Seeking speakers
In the same timeline
Consent for participating in the study
Consent for audio recording
Embedding of audio recording
Recording process
In the same timeline
Answering of Questionnaire
Recording of data
Data encoding
Analysis of attitudes towards pedagogical code-switching

Figure 4.1. Data Gathering Procedure

Data Analysis

For research question 1, mean and standard deviation was used. Mean is under the measures of central tendency that pertains to the average value of a group of numbers. On the other hand, the standard deviation is under the measures of variations which gives an idea of how much variation there is within a group of values. It aims to identify the measurement of the deviation (difference) from the mean (average) of a group (Skyes, et al., 2016) which will be associated with the dependent variables. For research question 2, the researchers examined the data collected using Friedman ANOVA and Kendall coefficient of concordance, wherein Friedman ANOVA is a nonparametric alternative to one-way repetition of measures analysis of variance (TIBCO, 2021) utilized to determine differences among groups when the dependent samples are being measured in an ordinal scale. Kendall coefficient of concordance is another nonparametric statistics that portrays the correlation between numerous cases (TIBCO, 2021). Utilizing the instruments, the researchers will accumulate qualitative data that are in ordinal level for the dependent variables, whereas the significant differences in the attitudes of the participants towards Tagalog, English, and CS, were identified.

5. Results

This section tackles the results gathered by the researchers from the participants.

5.1 Attitudes towards pedagogical English, Tagalog, and CS

Table 4 presents the means, standard deviation, and attitude interpretations derived from the verbal Guise test and analysis

Attitudes	English		Tagalog		CS	
	Mean	S.d.	Mean	S.d.	Mean	S.d.
Educated vs. Uneducated	3.45	0.66	3.60	0.60	3.27	0.73
Rich vs. Poor	3.02	0.76	3.15	0.74	2.88	0.78
Honest vs. Dishonest	3.37	0.66	3.56	0.60	3.35	0.69
Intelligent vs. Unintelligent	3.27	0.70	3.53	0.62	3.22	0.73
Sociable vs. Unsociable	3.28	0.76	3.29	0.73	3.27	0.80
Confident vs. Unconfident	3.44	0.75	3.37	0.73	3.18	0.83
Energetic vs. Lazy	3.23	0.73	2.97	0.74	2.90	0.86
Enthusiastic vs. Hesitant	3.21	0.75	3.15	0.70	3.02	0.83

Table 5.1. The Attitudes Towards English, Tagalog, and CS

Within the given data, the attitudes of the participants towards English, Tagalog, and Tagalog-English pedagogical CS are determined. The researchers used a 4-point Likert scale; hence, values ranging from 2.51 to 4.00 show a positive attitude toward the dependent variables. From the results gathered, the values in the mean columns lie within the positive range, indicating that the attitudes towards each language variable are positive adjectives. The attitudes towards Tagalog garnered the highest overall mean, proving that most participants favored the language's pedagogical use and significantly differed among all attitudes. In contrast, the attitudes toward CS resulted in a value of 3.14 for its mean average, which states that it is less positive than the pedagogical use of English but not higher than the attitude towards Tagalog. Conclusively, the data obtained for research question one is significant.

5.2 Differences in attitudes towards English, Tagalog, and pedagogical CS

*N= 784, *df*= 2, p= 0.00 except sociable vs. unsociable; p= 0.73

Attitudes	*df*	p	English		Tagalog		CS	
			S.d.	r	S.d.	r	S.d.	r
Educated vs. Uneducated	2	0.00	0.66	2.02	0.60	2.19	0.73	1.80
Rich vs. Poor	2	0.00	0.76	2.15	0.74	2.01	0.78	1.85
Honest vs. Dishonest	2	0.00	0.66	1.93	0.60	2.16	0.70	1.91
Intelligent vs. Unintelligent	2	0.00	0.70	1.90	0.62	2.23	0.73	1.87
Sociable vs. Unsociable	2	0.73	0.76	2.02	0.73	1.99	0.80	1.99
Confident vs. Unconfident	2	0.00	0.75	2.12	0.73	2.04	0.83	1.84
Energetic vs. Lazy	2	0.00	0.73	2.21	0.74	1.90	0.86	1.88
Enthusiastic vs. Hesitant	2	0.00	0.75	2.10	0.70	2.01	0.83	1.89

Table 5.2 Significant Differences towards English, Tagalog, and Pedagogical CS

In general, the researchers had observed that the findings show homogenous results within the responses of the participants among the three language types which are based on the stated standard deviations, due to the similar rankings evident in the evaluation of the participants.

Furthermore, the findings suggest that there are mostly significant differences in the attitudes towards Tagalog, English, and CS, with Tagalog having the most significantly positive attitude. While the data shows lower levels of positive attitudes towards CS, there is a significant level of difference for monolingual Tagalog and English languages.

5.3. Differences in Participants' Attitudes towards English, Tagalog, and CS

5.3.1. Educated vs. Uneducated

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded educated (positive attitudes) or uneducated (negative attitudes) $(df\,(2)\ =\ 159.20,\ p\ <\ 0.001)$. The box and whisker plot showed that the use of monolingual Tagalog and monolingual English has a significantly higher positive attitude when it comes to sounding educated compared to CS. Lastly, the use of CS does not mean sounding uneducated, but definitely sounding the least educated when compared to monolingual Tagalog and English.

5.3.2. Rich vs. Poor

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded rich (positive attitudes) or poor (negative attitudes) $(df\,(2)\ =\ 82.47,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that monolingual

Tagalog had the most significantly positive attitude of sounding rich than monolingual English and CS. Lastly, the use of CS does not mean sounding poor, but definitely sounding the least rich when compared to monolingual Tagalog and English.

5.3.3. Honest vs. Dishonest

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded honest (positive attitudes) or dishonest (negative attitudes) $(df\,(2)\ =\ 85.78,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that monolingual Tagalog had the most significantly positive attitude of sounding rich than monolingual English and CS.. Lastly, both the use of monolingual English and CS does not mean sounding dishonest, but only means sounding the least honest when compared to monolingual Tagalog.

5.3.4. Intelligent vs. Unintelligent

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded intelligent (positive attitudes) or unintelligent (negative attitudes) $(df\,(2)\ =\ 145.79,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that monolingual Tagalog had the most significantly positive attitude of sounding intelligent than monolingual English and CS. Lastly, both the use of monolingual English and CS does not mean sounding unintelligent, but only means sounding the least intelligent when compared to monolingual Tagalog.

5.3.5. Sociable vs. Unsociable

Using Friedman's test, the findings suggest that there is no significant difference on the evaluation of whether the speakers sounded sociable (positive attitudes) or unsociable (negative attitudes) $(df\,(2)\ 0.64,\ p\ >\ 0.001)$. The box and whisker plot clearly shows that when it comes to which language sounds more sociable, there is no significance of attitudes towards monolingual Tagalog, monolingual English, and CS. However, monolingual Tagalog gained a higher level of positive attitude when it comes to sounding sociable.

5.3.6. Confident vs. Unconfident

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded confident (positive attitudes) or unconfident (negative attitudes) $(df\,(2)\ =\ 74.13,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that the use of monolingual English and monolingual Tagalog has a significantly higher positive attitude when it comes to sounding confident compared to CS. Lastly, the use of CS does not mean sounding unconfident, but definitely sounding the least confident when compared to monolingual English and Tagalog.

5.3.7. Energetic vs. Lazy

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded energetic (positive attitudes) or lazy (negative attitudes) $(df\,(2)\ =\ 108.18,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that monolingual English had the most significantly positive attitude of sounding energetic than monolingual Tagalog and CS. Lastly, both the use of monolingual Tagalog and CS does not mean sounding lazy, but only mean sounding the least energetic when compared to monolingual Tagalog. The significant differences of the English language are due to its structure of words.

5.3.8. Enthusiastic vs. Hesitant

Using Friedman's test, the findings suggest that there is a significant difference on the evaluation of whether the speakers sounded enthusiastic (positive attitudes) or hesitant (negative attitudes) $(df\,(2)\ =\ 34.55,\ p\ <\ 0.001)$. The box and whisker plot clearly shows that monolingual English had the most significantly positive attitude of sounding enthusiastic than monolingual Tagalog and CS. Lastly, both the use of monolingual Tagalog and CS does not mean sounding hesitant, but only means sounding the least enthusiastic when compared to monolingual Tagalog.

6. Discussions

6.1. Attitudes towards pedagogical English, Tagalog, and CS

Participants' evaluations of speakers revealed that all three language types received positive ratings. Asuncion and Madrunio (2017) observed that students choose Tagalog since it is the country's primary language, meaning that it is the preferred language because it is spoken by everyone. Because Tagalog is primarily associated with national unity and linguistic identity (Gonzales, 1998), it was discovered that students perform better in the language with which they are most comfortable, indicating that Tagalog was the highest-rated language among the speakers. The English language has become the language of instruction in universities due to its ease of communication, allowing students and teachers to better grasp pedagogic interactions. Because it is commonly used by people of many mother tongues, it was employed to promote and bridge communication between students and teachers, resulting in a good attitude among the participants. Participants gave CS a positive evaluation because it improves learners' motivation in group involvement and shared meaning, indirectly but naturally monitors students' comprehension levels, and serves a useful function for teachers as well (Borlongan, 2012). The use of CS among most educational institutions in the Philippines proves its linguistic phenomenon that CS is a credible and effective tool for instructional practices that most Filipino universities are using in developing knowledge, understanding, and emphasis among learners.

6.2. Differences in attitudes towards English, Tagalog, and pedagogical CS

The participants show a positive attitude towards the three language types and that there is still a significant difference among them, wherein Tagalog received the highest level of positive attitude, followed by the English language, then followed by the CS. It is an unexpected finding that Tagalog is the most favorable language to be utilized in a pedagogical setting. This can be coincided by the previously said study of Asuncion and Madrunio (2017). In addition, in the study of Borlongan (2011), Filipino is still the preferred national language among pupils. It was also found in the study of Cahapay (2020) that the participants displayed a favorable attitude towards their mother tongue. On another note, despite being discovered in the study of Valerio (2015) that in a Philippine Context, learners express positive attitudes towards the English Language as their second language, they portrayed a more positive attitude towards their mother tongue as the language to be utilized when communicating in a pedagogical setting. Tagalog, being known as the widely spoken and the most easily understood language in all Philippine regions, received the highest positive attitude towards the speaker who used it in discussing a certain topic in communication when it comes to sounding rich.

English language is more favored than CS despite CS being largely prevalent among educated, middle-class, and upper-class urbanities in the Philippines due to the fact that Filipinos are not likely to abandon one language for the social benefits of another. For instance, When Lesada (2017) visited the Philippines, she discovered that a family in Cebu had decided to raise their young children solely in English because of the educational, socialization, and economic benefits it would provide. Because of the language's natural tone and structure, the differences in enthusiasm, confidence, and energy all demonstrated that English has the most positive meaning.

Despite the fact that many researchers, particularly in the fields of language and education, have suggested that CS can be used as a tool for effective teaching and learning, the evidence in this study suggests that CS is still less accepted in pedagogical settings than English and Tagalog. This is supported by Bautista's (2004) research, which found that after the bilingual policy was announced, debates arose about whether Taglish CS should be allowed as a mode of communication in academic institutions, particularly for the transmission of classroom instructions and academic discourse in general. Furthermore, according to several researchers, CS does not inspire passion among instructors, and it was explained that CS is associated with ineptitude and interference (Bautista, 2004; Flores, 2019; Goulet, 1971; Labitigan, 2013; and Johansson, 2013). The levels of positive attitude towards the three language types were all similarly sociable due to CS, just like English and Tagalog, is used to build closer relationships between students and their teachers. And lastly, the participants showed a lower level of positive attitude towards CS due to the speaker monitoring more the construction of every word and grammar of their sentences when using CS, compared to naturally speaking a monolingual language (Coor, 2019).

7. Conclusions

It was discovered that the students have a positive attitude towards all three language types in a pedagogical context. However, there is an evident difference among the participants' attitudes wherein

Tagalog garnered the most positive attitude, followed by English and lastly CS.

This study suggests that despite CS being acknowledged as an effective medium in a pedagogical setting, CS is still less favored by the students when it comes to teaching and learning, being manifested as the least educated, rich, honest, intelligent, sociable, confident, energetic, and enthusiastic. Furthermore, the study found that regardless of the social benefits the English language could bring to a student, Tagalog received the highest level of positive attitude from the participants, ranking first in sounding educated, rich, honest, intelligent, and sociable. The verbal Guise method was used to gather data in this study, with students being asked to listen to and then evaluate the speaker, which can be considered an innovative and more natural means of examining the participants' attitudes. As a result, this research can contribute to language attitude research in the pedagogical context, which is gaining popularity at the moment.

8. Recommendations

As for the implications for research, the study suggested a positive attitude towards English, Tagalog, and CS, with CS receiving the least positive one in a pedagogical set-up. Another way of collecting data can be used to conclude a more accurate result and further understand the reasons behind the attitudes towards each language type.

A four-point Likert scale in the verbal guise test identifies the different attitudes, which may be positive or negative. This is deemed to be used to collect extreme feedback from the respondents without providing a neutral option, resulting in a more specific response. This type of scale is ideal and can be evenly split into two divisions, top two choices and bottom two choices, as said by Hopper (2016), making reporting of results easier and simpler. The researchers also suggest that future studies about language attitudes should solely focus on the Tagalog language, as it only has a limited study. Since this study showed a positive evaluation among the variables, future researchers may also examine if monolingual English and monolingual Tagalog will receive a negative attitude if studied individually. Furthermore, future researchers may also choose to use a different subject and topic concerning what the speakers in the verbal guise test will use.

As for the implications for practice, although all three language types were deemed to have received a positive response, the use of Tagalog received the highest positive attitudes from the respondents. With this given, the implementation of the DepEd's MTB-MLE policy can be strengthened and put into practice as this study is an evidence that Filipino students express the most positive attitude towards Tagalog. Moreover, if the use of English persists, activities and strategies should be done to improve students' attitudes towards the English language in a school set-up. Through this, the most effective language type that can be used in pedagogy will be deemed to benefit both the students and teachers.

References

Abad, L. (2005). Codeswitching in the Classroom: A Clash of two languages. Miriam College Faculty Research Journal, Vol. 25, pp. 36-52.

Abad, L. (2010). Code-switching: An Alternative Resource in Teaching Science and Math. Miriam College Faculty Research Journal, Vol. 32, No. 1.

Adeeb, P. M. (1994). A Quasi-Experimental design to study the effect of multicultural coursework and culturally diverse field placements on preservice teachers' attitudes toward diversity. *UNF Graduate Theses and Dissertations*.

Alieto, E. (2018). Language shift from English to mother tongue: Exploring language attitude and willingness to teach among pre-service teachers. *TESOL International Journal*, *13*(3).

Asuncion, Z. S., & Rañosa-Madrunio, Ph.D., M. (2017). Language Attitudes of the Gaddang Speakers towards Gaddang, Ilocano, Tagalog and English. *Studies in English Language Teaching*, *5*(4). https://doi.org/10.22158/selt.v5n4p720

Barnes, B. R. (2019). Quasi-experimental designs in applied behavioural health research. *Transforming Research Methods in the Social Sciences*, 84–96. https://doi.org/10.18772/22019032750.11

Bautista, M. L. (1991). Codeswitching studies in the Philippines. International Journal of the Sociology of Language, 88, pp. 19-32.

Bautista, M. L. (1999). An analysis of the functions of Tagalog-English Code-switching: data from one case. The Filipino Bilingual: A Multidisciplinary perspective (Bautista, M. L. & Tan, G., Eds.). Manila: Linguistics Society of the Philippines.

Bautista, M. L. (2004). Tagalog-English code switching as a mode of discourse. Asia Pacific Education Review. pp. 226-233.

Bolton, K. and Bautista, M.L. (2004). Philippine English: tensions and transitions. World Englishes, Vol. 23, No. 1, pp. 1-5.

Borlongan, A. (2009). Tagalog-English Code-switching in English Language Classes: Frequency and Forms. TESOL Journal, Vol. 1. Pp. 28-42.

Campbell, D., Cook, T., & Shadish, W. (2002). Experimental and quasi-experimental designs for generalized causal inference. *Boston: Houghton Mifflin.* 1-504, https://www.alnap.org/system/files/content/resource/files/main/147.pdf

Chan, J. Y. H. (2018). Gender and attitudes towards English varieties: Implications for teaching English as a global language. System, 76, 62–79.

Choi, N., Cho, H. J., Kang, S., & Ahn, H. (2021). Korean Children's Attitudes toward Varieties of English:The Role of Age and English Learning Environment. Languages, 6(3), 133. https://doi.org/10.3390/languages6030133

Gonzales, A. (2004). The social dimensions of Philippine English. World Englishes, Vol. 23. No. 1, pp. 7-16.

Gumperz, J. J. (1982). Conversational Codeswitching. In Discourse Strategies. Cambridge, England: Cambridge University Press. International Journal of Education and Research Vol. 4 No. 11 November 2016 45

Language Attitude towards Using Code-switching as a Medium of Instruction in the Case of Wollo University Kombolcha Institute of Technology, Ethiopia. (2019). *Research on Humanities and Social Sciences*. Published. https://doi.org/10.7176/rhss/9-21-04

Lund Research Ltd. (2018). *Friedman Test in SPSS Statistics - How to run the procedure, understand the output using a relevant example | Laerd Statistics*. Laerd Statistics. https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php

Mahato, R. (2016). Code-switching in classroom teaching. *Academic Voices: A Multidisciplinary Journal*, *6*(1), 28–31.Mechatronics Engineers' Perception of Code Mixing: Philadelphia University and Hashemite University as a Case Study. (2016). *International Journal of Applied Linguistics and English Literature*, *5*(7). https://doi.org/10.7575/aiac.ijalel.v.5n.7p.110

Natividad, M. R. A., & Batang, B. L. (2018). Students' perceptual learning styles and attitudes toward communicative language teaching. *TESOL International Journal*, *13*(4).

Nejjari, W., Gerritsen, M., van Hout, R., & Planken, B. (2019). Refinement of the matched-guise technique for the study of the effect of non-native accents compared to native accents. Lingua, 219, 90–105.

Nurhamidah, N., Fauziati, E., & Supriyadi, S. (2018). Code-switching in efl classroom: Is it good or bad? *Journal of English Education*, *3*(2), 78–88.

Nyoman Wartinah, N., & Wattimury, C. N. (2018). Code switching and code mixing in english language studies' speech community: A sociolinguistic approach. In *Berumpun journal*, *1*(1).

Phan, H. L. T. (2020). VIETNAMESE LEARNERS' ATTITUDES TOWARDS AMERICAN AND BRITISH ACCENTS. *European Journal of English Language Teaching*, *6*(2). https://doi.org/10.46827/ejel.v6i2.3498

Price, P. (2015). *Research methods in psychology, 2nd canadian edition*. BCcampus.

Reid, L. A. (2008). Who are the indigenous? origins and transformations. 1st International Conference on Cordillera Studies. UP, Baguio City. February 7-9, 2008

Rivera, J. (2015). A Quasi-Experimental study on the impact of explicit instruction of science text structures on Eighth-Grade english learners' and Non-English learners' content learning and reading comprehension in three inclusive science classrooms. *Electronic Theses and Dissertations, 1245*. https://stars.library.ucf.edu/etd/1245

Rose, S. & Dulm, O. (2006). Functions of code switching in multilingual classrooms. *Per Linguam, 22*(2), 1-13.https://doi.org/10.5785/22-2-63

Salmon, W. (2017). Language attitudes, generations, and identity in coastal Belize. *African and Black Diaspora: An International Journal*, *10*(3), 299–312. https://doi.org/10.1080/17528631.2017.1300212

Salmon, W., & Gómez Menjivar, J. (2016). Language variation and dimensions of prestige in Belizean Kriol. *Journal of Pidgin and Creole Languages*, *31*(2), 316–360. https://doi.org/10.1075/jpcl.31.2.04sal

Sun, G. B., & Eun, H. H. (2005). A Quasi-Experimental research on the educational value of performance assessment. *Asia Pacific Education Review*, *6*(2), 170–190.

Tahir, A., Fatima, I., & AbuzAar, N. (2016). Teachers' and students' attitude toward code alteration in pakistan english classrooms. *Journal of English Education and Linguistics Studies, 3*(1), 85-108.

TIBCO. (2021). *Nonparametric statistics notes - friedman ANOVA and kendall concordance*. TIBCO Software Inc. https://docs.tibco.com/pub/stat/14.0.0/doc/html/UsersGuide/GUID-2F86358E-28E3-45C1-98D9-4257CAB12073.html

Thompson, R. (2003). Filipino, English, and Taglish: language switching from multiple perspectives. Philadelphia: John Benjamins Publishing Corporation

Valerio, T. B. (2015). Filipino – English Code Switching Attitudes and Practices and Their Relationship to English Academic Performance among Freshman Students of Quirino State University. *International Journal of English Language Teaching*, *2*(1).

Syllabified Sequence-to-sequence attention for transliteration

G.Vyshnavi
vgutta7@gatech.edu

G.Sridevi
dr1sridevi@vrsiddhartha.ac.in

M.Ritesh
ritesh.s.m@ril.com

P. Krishna Reddy
pkreddy@iiit.ac.in

Abstract

The problem of transliteration deals with the phonetic transcription of text from a source writing system into a target writing system. With the inception of neural net models like Sequence-to-sequence networks, transliteration has seen significant progress in the last decade. However, the accuracy of such systems is still far from ideal. This is made more appar- ent when the source text to be transliterated itself is entered incorrectly by intermediary users, further degrading the performance. In this paper, we propose Syllabified Sequence- to-Sequence network (Syll-S2S) towards im- proving the transliteration quality from Roman script to low-resource Indic scripts like Devana- gari. As part of this, we present the rules of Sonority sequencing principle to Devanagari and other Indic scripts. In addition, the proposed framework incorporates Elastic Search stack which maps incorrect transliterations to their existing reference transliterations for handling erroneous entries of source texts. Experi- ments demonstrate significant performance im- provement of the proposed framework with re- spect to the existing schemes.

1 Introduction

English is one of the most widespread foreign languages in India, a home to 22 official languages and more than 1000 dialects written in more than 14 different scripts. With the rapid advancements in worldwideweb and mobile devices, people in India create, share, tag and search multifaceted data multi- lingually but mostly using the Roman or Latin script (Chanda et al., 2010) across different mediums. Thetext written in a native language, but using a non- native script like Roman, mostly does not follow any standard spelling rule, but uses the orthography of the script based on pronunciation of the words. This process of phonetically transcribing a word or text from one writing system into the another writing system such that the pronunciation of the word remains same is called Transliteration.

Transliteration is a part of Natural Language Processing (NLP) and has several useful applications; Cross language information retrieval, Machine translation etc. It has wide ramifications in low-resource languages in general, where web presence is limited, specifically for Indian languages. A substantial por- tion of textual data being generated and queried upon the web belongs to the transliteration domain, thus containing a good amount of information and therefore needs to be studied.

In this paper, we focus on transliteration from the most commonly used Roman script to native Indian scripts. Most of the major Indian language scripts are derived from the ancient Brahmi script and consequently are highly phonetic in nature. Hence, we primarily focus on the task of transliteration from Roman to Hindi. Hindi, an Indo-Aryan language, written in Devanagari, is the lingua-franca of India. We therefore consider Hindi as the primary link for Roman to native Indian script transliteration, as the quantity of Hindi literature (especially online) is more than twice as in any other Indian language.

For the Roman-Hindi transliteration task, the earliest methods are rule based approaches (Goyal and Lehal, 2009; Kang and Kim, 2000; Jia et al., 2009) which involved character, phoneme and grapheme matching between the parallel transliteration corpora. But the rule based approaches fall short due to the several exceptions possible. Another popular method is Anoop et al.'s Indic-nlp (Kunchukuttan et al., 2020), which is a statistical machine translation based approach relying on language model training. Recently, Sequence-to-sequence attention based networks (seq-to-seq) proposed in (Sutskever et al., 2014), have garnered wide attention for machine translation (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017) and works (ud Din, 2019; Ameur et al., 2017; Mandal and Nanmaran, 2018) extended the same for the transliteration task. However, the existing seq-to-seq models too have failed in producing desirable results.

In this paper, we develop a Syllabified Sequence-to-sequence net (Syll-S2S) for improving the quality of Roman to Devanagari transliteration w.r.t. the state-of-the-art works. The proposed approach uses Sonority Sequencing Principle (SSP) to get the syl- lables of the source (and target data during training) thus enabling the enforcement of syllable-syllable at- tention. This way, the model would be able to learn the target data with a greater precision and would also enable faster knowledge acquisition than con- ventional seq-to-seq models.

Additionally, we consider an important practical aspect with regards to transliteration which is the user input. In several applications, the primary source of the input text in Roman is derived from an intermediary user. In such scenarios, the user given input word may have slight variation from its correct form in the supposed cases of well-established words like dictionary-based-words or named entities. For handling such cases, we incorporate Elastic Search stack (ES stack); a distributed, open source search and analytics engine (Gormley and Tong, 2015). In the case where reference transliterations are available, we use fuzzy search query (Gu et al., 2018) on ES stack with added constraints on consonant match for mapping the model output transliterations and the reference transliteration.

The main contributions of this paper are 3-fold.

1. We develop a novel Syllabified Sequence-to-sequence model (Syll-S2S) for improving the transliteration quality from Roman to Devanagari in comparison to the state-of-the-art.

2. In the case where true transliterations in Devanagari are available, we incorporate fuzzy searching on user-given Roman input for accurate matching to its indexed reference transliteration with Elastic Search stack.

3. We have demonstrated the superiority of the proposed approach with extensive performance evaluation.

The remainder of this paper is organized as follows. In Section 2, we present the related work. We present the background of seq-to-seq net, ES stack and inverse mapping in Section 3 followed by proposed approach in Section 4. Experimental results and conclusions are presented in Section 5 and Section 6 respectively.

2 Related work

In this section, we briefly describe few of the existing works which do transliteration.

Methods in (Goyal and Lehal, 2009; Kang and Kim, 2000; Jia et al., 2009) incorporate rule-based transliteration which can again be divided into 3 categories. (1) Character mapping approach (Goyal and Lehal, 2009) uses character mapping for doing transliteration. Under this approach, the characters of source script are mapped to those of the target script on the basis of pronunciation. Character mapping does not give very good results as the pronunciation of characters and the total number of character varies from script to script. (Kang and Kim, 2000) uses (2) Phoneme Based Approach which defines the relation and correspondence between the phonemes of the source and target script. An alignment of the phoneme for the characters of source script to the phoneme of the target script is done using methods like language modeling (Chelba and Jelinek, 2000). (Jia et al., 2009) uses Phoneme Based Approach by defining the relation and correspondence between the (3) graphemes of the source and target scripts.

The second class of works is based on statisti- cal machine translation (SMT). Anoop kunchkuttan et al. propose Indic-NLP in (Kunchukuttan et al., 2020) and develop a SMT approach with a Moses de- coder for transliteration. Moses (Koehn et al., 2007) allows us to automatically train translation models for any language pair. It uses phrase based transla- tion Models and word alignments.

With the recent advancements of Sequence-to-sequence models (seq-to-seq), transliteration quality has improved greatly. Proposed in (Sutskever et al.,2014), these networks are used for translation of an input text from one language to another. Works (ud Din, 2019; Ameur et al., 2017; Mandal and Nan- maran, 2018) have later extended seq-to-seq mod- els for transliteration as well and various attention- based seq-to-seq models have been subsequently been proposed (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017). Attention based seq-to-seq models (Bahdanau (Bahdanau et al., 2014)) work by using an encoder to learn representations of the input sequence and a decoder to produce the out- put sequence from the hidden representations the encoder created. Few attention variants in seq-to- seq architectures include (Luong et al., 2015) by Luong et al. and self-attention (Vaswani et al., 2017). Luong attention differs from Bahdanau in the alignment calculation and the position at which the attention mechanism is introduced in the decoder. Self-attention introduced in (Vaswani et al., 2017), is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence.

3 Background

In this section, we explain the sequence-to-sequence networks and present the details of our employed Elastic Search stack.

3.1 Sequence-to-sequence network

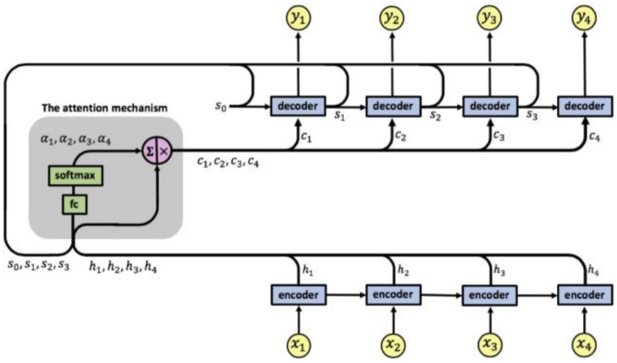


Figure 1: Sequence-to-sequence network

Sequence-to-sequence network (seq-to-seq) (Sutskever et al., 2014) is used for converting one sequence into another, particularly when the input and the output sequence lengths vary. They use encoder-decoder networks (Figure 1). Their details are presented below.

1. Encoder: The encoder has sequential recurrent layers which learn to encode the input data accurately and produce a set of hidden states which are passed to the decoder.

2. Decoder: The decoder takes the states from encoder and uses it to generate context vector at every time-step t. Context vector holds the weighted cumulative information from all of encoder hidden states and varies across the time-steps. At every time-step, the previous decoder hidden state along with the corresponding context vector is passed as input to the recurrent layer at t. The joint output from all time-steps gives us the output text sequences.

The above network employs attention so that the decoder learns to focus on relevant encoder hidden states. It mainly employs two kinds of attention,

- Global attention: considers all the hidden states in creating the context vector.

- Local attention: considers only a subset of the hidden states in creating the context vector.

We train all our seq-to-seq methods using Adam optimiser. The encoder and decoder have bi-directional GRU cells (Deng et al., 2019).

3.2 Elastic Search stack

Consider cross language information retrieval, a well known user-driven application. Such an application consists of transliterating and matching a user-given input in Roman to its already indexed reference transliteration in the target domain (Eg: Devanagari). In such cases, the transliteration framework's output should have to match accurately to the source, regardless of the slight variations in the user-given input in Roman script from its originally intended form. To handle this, we have built an end-to-end framework which incorporates Elastic Search stack into the proposed framework for precise matching to the reference after transliteration. Elastic Search stack (ES stack) (Gormley and Tong, 2015), is a distributed, open source search and an analytics engine. It can also easily map rogue model predic- tions to their references.

3.2.1 Terms

- Index: Adding 'data' to ES stack is known as "indexing." In our case, we can either index thetrue scripts of native words to Devanagari or theknown transliterations of non-native words.

- Mapping: It is the process of defining how a document, and the fields it contains, are stored and indexed.

- Fields: Fields are properties in a mapping. Every mapping contains a list of fields or properties pertinent to the document. In our case we use mappings to define the properties of the words being indexed. There are 2 types of fields: *Key- word* and *text*. Keyword fields are only search- able by their exact value. Text field allows search for individual words within each full textfield.

- Analyzer: The analyzer parameter specifies theanalyzer used for text analysis when indexing or searching a text field. The default analyzer for *Keyword* type is standard and is immutable.

- Match and Multi-match query: A match query returns documents that match a provided text which is further analyzed before matching. The multi-match query builds on the match query to allow multi-field queries. We can influence scoring as needed by prioritizing more importantfields.

- Fuzziness parameter: Adding fuzziness parameter [1] to a multi-match query turns a plain multi-match query into a fuzzy one. It generates matching terms that are within the maximum edit distance specified in the parameter and then checks the term dictionary to find out which of those generated terms actually exist inthe index.

3.2.2 Inverse mapping with fuzzy search

We have used 3 fields in the mapping we have built.Their details are presented below.

1. *Devanagari_script_field*, denotes the script ofa word being indexed in Devanagari. The field type is *keyword*, so the search on this field happens full-text.

2. *Roman_script_field*, denotes the script of the word being indexed in Roman. The field type is*keyword*.

3. *Consonants_field*, denotes the sequence of just the consonants of the word being indexed. The field type is *text*, so the search on this field hap-pens consonant wise. For this, we use an icu an- alyzer [2] with icu tokenizer [3] and a custom char filter which converts all the vowels in the text being analyzed to NULL using the predefined unicode mappings of vowels in Devanagari.

The search works as follows. The transliterated can- didates are queried against the ES stack consisting of the indexed reference transliterations. This is done using a multi-match query which jointly queries on the fields *Devanagari_script_field* and *Conso- nants_field* with a fuzziness score of 0.7. Those can- didates which are mapped to a reference transliter- ation are updated to be the same as the reference before being returned.

4 Proposed framework: Syll-S2S

In this section, we propose a novel Syllabified Sequence-to-sequence model for improving the performance of Roman to Devanagari transliteration to that in the existing approaches. We first present the basic idea of the proposed approach. Next, we explain the proposed approach in detail.

4.1 Basic idea

Local attention based sequence-to-sequence models have been under-explored for the task of transliteration and could dramatically improve the knowledge gain with respect to global attention based sequence-to-sequence models. Leveraging this, the idea is to segment both the input Roman text and its parallel counterpart Devanagari text into syllables before building the attention based seq-to-seq model. Syllable is a unit of spoken language consisting ofa single uninterrupted sound formed generally by a vowel (Eg: a,e,i,o,u) and preceded or followed by one or more consonants (Eg: b,c,d,..). By reducing the sequences into their constituent syllables offline, we can employ a teacher forcing method and force the model to attend to a fixed window length of sylla- bles at each decoding step in the attention part of decoder in seq-to-seq networks. At each decoding time-step, a window centered around the source po- sition based on the syllable alignment is used to com- pute the context vector for the syllable corresponding to the target position. Thus, the attention weights at each time-step are distributed and limited to the corresponding syllables of the positions in the win- dow which allows effective knowledge building by the model.

4.2 Sonority Syllabification

The process of splitting a text into its constituent syllables is referred to as Syllabification (Treiman and Zukowski, 1990). Since it is derived directly from pronunciation, syllables are script-agnostic and are vital information pillars for machine translation and transliteration tasks. For doing syllabification, we utilise Sonority syllabification principles (SSP) available for Roman script (Henke et al., 2012). The Sonority Sequencing Principle (SSP) or Sonority se- quencing constraint is a phonetic principle that aims to outline the structure of a syllable in terms of sonority. Basing on the SSP rules for Roman, we present the SSP rules for Devanagari and other In- dic scripts. We are the first to present SSP rules for Indic scripts.

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-fuzzy-query.html

[2] https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-icu.html

[3] https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-icu-tokenizer.html

4.3 Syllabified local attention with seq-to-seq net

Consider the Roman text ‘Hajagiree’ and its reference transliteration `हजिगरी'. Observe the one-to-one mapping between the source (Roman) and the target (Devanagari) text syllables (Figure 4.3(a)). Moreover, this mapping is both static and sequential. Based on this, we propose Syllabified Sequence-to-sequence model using local attention. Recall from Section 3.1 that in a local attention based seq-to-seq, at a given time-step, the decoder is fed only a part of encoder context information. This information is generally limited to the corresponding input part.

With syllabification, the decoder needs to attend only on the select parts of the encoded input data specified by us at each time-step. This way, the model can learn to decode more accurately when it focuses on just the needed input syllables rather than the parts which it does not depend on. Thus, the attention scope is limited to just the input syllable/s corresponding to the present decoder time-step. However, considering an x-length window towards left and right could prove beneficial as syllabification is not always perfect (Figure 4.3(b)).

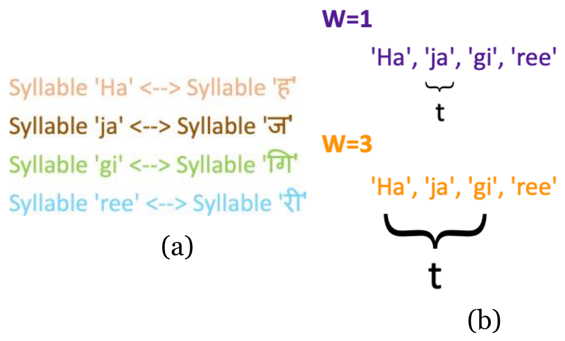

Syllable 'Ha' <--> Syllable 'ह'
Syllable 'ja' <--> Syllable 'ज'
Syllable 'gi' <--> Syllable 'गि'
Syllable 'ree' <--> Syllable 'री'
(a)
W=1
'Ha', 'ja', 'gi', 'ree'
t
W=3
'Ha', 'ja', 'gi', 'ree'
t
(b)

Figure 2: (a) shows the syllable-syllable correspondence between an example’s input and output texts. (b) shows the attention windows W={1,3} focused by the decoder when at t^{th} time-step

The pseudo-code of the test process is as follows.

1. Input text in Roman script is cleaned and syllabified using SSP principles.
2. The preprocessed input text is sent to the trained Syll-S2S for transliteration to one of the Indic scripts. Outputs Top-N (N=3) candidate transliterations.
3. The transliterations are queried using inverse mapping described in Section 3.2.2 and the mapped candidate transliterations are updated to their references.
4. Returns the final candidate transliterations.

5 Experimental setup

All the experiments are conducted on an Intel i5 processor with 8GB RAM running Ubuntu Linux oper-ating system.

5.1 Dataset

It is to be noted that the parallel transliteration corpus i.e, the text in Roman and its counter transliteration in Devanagari is very limited and had to be scraped from several existing open sourced works. We have 1.5 million parallel words worth of data in Roman and Devanagari after scraping. The words in the resulting data are diverse ranging from named entities to native words from Roman and Devana- gari. We use a train-val-test split of 4:1:1 and train all the neural net models for 100 epochs.

5.2 Data preprocessing

Data cleaning and preprocessing was done on both the corpus before model building. This includes stripping unwanted characters like tags, HTML entities, UNK tokens, special characters etc and lower-casing the Roman data corpus. We use UTF-8 en- coding on both the source and target corpus.

Once cleaning is done, for implementing the proposed approach, we pass the input text to the Sonority syllabification module (Section 4.2) for splitting into syllables. The syllabified text is then passed as input to Syll-S2S model.

5.3 Methods and metrics

We have considered the following methods for performance evaluation.

- Rule-based (RL): Rule based transliteration system based on predefined character-to-character mapping.
- Google Transliterate (Google): Google’s officialopen API for transliteration [4].
- Indic-NLP (Indic) (Kunchukuttan et al., 2020).
- Seq-to-seq (S2S): A sequence-to-sequence model incorporating Luong’s attention based scoring (Luong et al., 2015).
- Self attention (SA): A self attention model based on transformers (Vaswani et al., 2017).

[4] https://inputtools.google.com/request?itc=hi-t-i0-und&num=4&cp=0&cs=0&ie=utf-8&oe=utf-8&app=demopage&text=’’

Scores	RL	Indic	S2S	SA	Google	Syll-S2S-W{1,3,5}		
						W1	W3	W5
Top-1 Accuracy	0.45	0.57	0.72	0.79	0.83	0.82	0.86	0.85
Top-2 Accuracy	0.52	0.66	0.75	0.81	0.84	0.84	0.87	0.87
Top-3 Accuracy	0.58	0.71	0.77	0.83	0.84	0.84	0.87	0.87
Levenshtein distance	0.63	0.77	0.85	0.90	0.92	0.92	0.94	0.94
BLEU before inverse mapping	0.39	0.48	0.63	0.71	0.74	0.76	0.77	0.75
BLEU after inverse mapping	0.46	0.59	0.75	0.82	0.85	0.85	0.87	0.87

Table 1: Performance comparison of the considered methods in all the metrics. The top scores are highlighted in redand the second best are highlighted in green.

- Syllabified monotonic Sequence-to-Sequence (Syll-S2S): The proposed approach involving monotonic attention from source to target syllables using S2S model. We also vary the syllables' window-size W={1,3,5} and the vari- ants are named as Syll-S2S-W1, Syll-S2S-W3, Syll-S2S-W5 respectively.

A beam size of 3 (means 3 candidate transliterations for every input) is employed for all the neural net methods (Freitag and Al-Onaizan, 2017).

The following performance metrics have been employed.

- Top-N accuracy: The average number of correct transliterations within the Top-N (N={1,2,3}) beam-search candidate transliterations of the source text.
- Levenshtein distance: The number of single-character edits required to change predicted transliteration into the correct reference.
- 3-gram match: The average number of word-level 3-grams with matches to their corresponding 3-grams from the reference text.
- BLEU scores before inverse mapping: The sim-ilarity of transliterated text to its reference be- fore inverse mapping.
- BLEU scores post inverse mapping: Fuzzy-search based inverse mapping (Section 3.2.2) is used to map the candidate transliterations to the existing reference transliterations in the Elastic Search stack (Section 3.2) before computing BLEU scores.

5.4 Results

5.4.1 Performance comparison

Table 1 shows the performance of all the implemented methods in various score metrics (leftmost column). The best scores highlighted in red are by the proposed approach Syll-S2S. Note the performance improvement from S2S to Syll-S2S. This is because of the monotonic local attention employed in the latter by enforcing syllable-syllable correspondence between the source and the target texts. In particular, Syll-S2S-W3 which is Syll-S2S with window-size 3 is performing the best overall with re- spect to the second best method Syll-S2S-W5 (win- dow size=5) highlighted in green. This is because of the reduced attention window size in Syll-S2S- W3 w.r.t. W=5 as in the former, only the imme- diate left and right syllables apart from the current syllables are focused while computing the attentionscores whereas in W=5, the left and the right window span increases by a step more leading to distributed attention. Between Syll-S2S-W1 and Syll-S2S-W3, the latter does well as vowels might be split between the current and the next/before syllables making the learning process more reliable when the attention window covers them beside just the current syllable. The model closely following the Syll-S2S-W{1,3,5} variants is the Google transliterate. Even the model with the best architecture, SA is still behind the pro- posed approach because of over-fitting due to limited data.

Effect of inverse mapping: It can be observed from the table that the BLEU scores have noticeably improved after incorporating fuzzy-search based in- verse mapping for all the approaches (Section 3.2.2). Recall that the above mapping matches the candi- date transliterations to their existing source translit- erations using fuzzy searching on consonants. The values depict the merit of mapping the candidate transliterations as most often there are only minor differences between the candidate and the reference transliterations, evidenced by higher values in the metric Levenshtein distance.

5.4.2 Effect of syllabified local attention

To understand the advantages of using syllabified monotonic local attention as in the proposed ap-

Sentence	Reference transliteration	Syll-S2S-W5 result	Syll-S2S-W3 result
My Birthday Song	माय बथर्डे संॉग	माइ िबरटहडाय सोंग	माय बथर्डे संॉग
Happy Phirr Bhag Jayegi	हैप्पी िफर भाग जायगी	हापपय िफर भाग जायेगी	हैप्पी िफर भाग जायगी
Wo India Ka Shakespeare	वो इंिडया का शेक्सपीयर	वो इंिडया का शक् ेस्पोरे	वो इंिडया का शेक्सपीयर
Yamla Pagla Deewana Phir Se	यमला पगला दीवाना िफर से	यमल पगला दीवानन िफरसे	यमल पगला दीवाना िफर से
Kaashi in the search of Ganga	काशी इन सचर् ऑफ़ गंगा	काशी इन सचर् ऑफ गंगा	काशी इन सचर् ऑफ़ गंगा
Mausam Ikrar Ke Do Pal Pyar Ke	मौसम इकरार के दो पल प्यार के	मौसम इकरार के दो पाल प्यार के	मौसम इकरार के दो पल प्यार के
Jal bin machhli nritya bin bijli	जल िबन मछली नृत्य िबन िब-जली	जळ िबन मछली नृत्य िबन िबजली	जाल िबन मछली नृत्य िबन िबजली
Albert pinto ko gussa kyu aata hai	अल्बटर् िंपटो को गुस्सा क्यों आता है	अल्बटर् िंपटो को गुस्सा क्यों आता ही	अलबटर् िंपटो को गुस्सा क्यों आता है

Table 2: Example Roman input sentences, their reference transliterations and the output transliterations from Syll- S2S-W5 and Syll-S2S-W3 methods. Results in red indicate exact match, in green indicate the results with edit distance=1 and in blue indicate the results with edit distance > 1 from the reference.

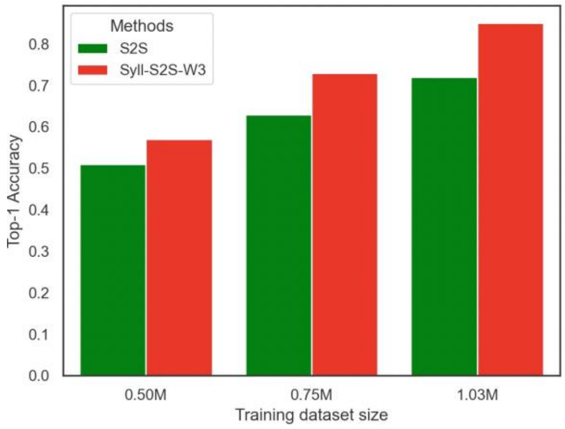


Figure 3: Top-1 Accuracy of S2S and Syll-S2S-W3 meth- ods on varying training data size

proach Syllabified Sequence-to-sequence (Syll-S2S) rather than global-attention as in conventional seq-to-seq (S2S) models, we compare Syll-S2S-W3 model (the best performer, see table 1) with S2S model at varying training data-set sizes. i.e., we train the two models by using the training sizes of 0.5,0.75 and 1 million before testing. Fig 3 shows the Top-1 accuracy vs training dataset-size results for the two approaches. As can be seen, with increasing size of training dataset, the rate of improvement for Syll-S2S-W3 is significantly greater w.r.t. S2S. This means that the proposed approach has a faster knowledge-acquisition rate than S2S model owing to the local attention in Syll-S2S from enforcing monotonic source-target syllable-wise attention. In conclusion, we can deduce that the accuracy improvement for the proposed approach as more training data resources become available would be substantial when compared to conventional seq-to-seq models.

5.4.3 Effect of attention window

To demonstrate the importance of the size of attention window in the proposed approach, we compare Syll-S2S-W3 (window size=3) with Syll-S2S-W5 (window size=5) (refer Figure 4.3(b)). Table 2 shows a qualitative comparison of results from both the methods on few example input sentences of varyinglengths. Results in red indicate exact match to ref- erence transliteration. Results highlighted in green indicate the results with edit distance equal to 1 from reference and ones in blue indicate the results with edit distance greater than 1.

As can be seen, between the two, Syll-S2S-W3 is clearly the best performer as it has the most predic-tions in red and green as compared to Syll-S2S-W5 which has more in blue. This is because of the re- duced attention window size in Syll-S2S-W3 w.r.t. W=3 as in the former, only the immediate left and right syllables apart from the current syllables are fo- cused while computing the attention scores whereas in W=5, the left and the right window span increases by one more step leading to a more scattered attention. This results in the two left and right syllables on both sides of the current syllable receiving around the same attention as the syllable to be decoded.

5.4.4 Comparison with Google transliterate

- -

Input 1: Jack aur jill pani ki ek balti lene keliye pahadhi par chad gaye
Reference: जैक और िजल पानी कɃएक बाल्टी लेने के िलए पहाड़ी पर चढ़ गए

Syll-S2S-W3 result: जैक और िजल पानी कɃ एकबाल्टी लेने के िलए पहाड़ी पर चढ़ गए
3-gram match score: 1

Google result: जैक और िजल पाि न कɃएक बाल्टी लेने केे िलये े पहाड़ ी पर चै ड गया
3-gram match score: 0.92

- -

Input 2: Uppu kappurambu nokka polika nundu chooda chooda ruchulu jaada veru purushulandu punya purushulu veraya viswadhaabhiraama, vinura vema
Reference: उप्पु कप्पूरबु नोक्का पोिलका नन्दू चूड चूड रुचुलु जाद वेरय पुरुशूलंदु पुण्य पुरुशुलु वेरय िवस्ववा"भरामा

Syll-S2S result: उप्पु कप्पूरबु नोक्क पोिलक नन्द चूड चूड रुचुलु जाद वेरया पुरुशूलंदु पुण्य पुरुशुलु वेरया िवस्वद्य ा"भरामा िवनुर वेमा
3-gram match score: 0.84

Google result: उप्पू कप्पूरबु नोक्का पोिलका नन्दू चूडा चूडा रुचुलु जादा वेरु पुरुशूलंदु पुण्य पुरुशुलु वेरय
िवस्ववा"भरामा
िवनुरा वेमा
3-gram match score: 0.71

- -

The above inputs depict two example sentences and their reference and Top-1 candidate transliteration from Syll-S2S-W3 (the best performer, see table 1) and Google transliterate. We also show the 3-gram match score which is the the average number of word-level 3-grams which are correctly matched to their corresponding 3-grams from the reference text. 3-gram match scores act as a direct measure of a model's strength when plugged in other crucial downstream tasks like Information retrieval, attribute linking etc and having a high 3-gram scores would be the ideal for all models.

There is a clear distinction in the scores from the two approaches with Syll-S2S-W3 giving higher val- ues in comparison to Google transliterate. Observe how the difference in scores becomes more apparent with increase in the input-length. This shows that in applications involving more input complexity, the proposed approach is expected to produce the best result.

6 Conclusions

In this paper, we have proposed Syllabified sequence-to-sequence attention model for improving the transliteration quality with respect to the current works. For this, we present the Sonority syllabification principles for Devanagari and other Indic scripts. To handle the erroneous entry of user-given text and to further boost the performance, we incorporate fuzzy-search based inverse mapping with consonants by employing Elastic Search stack. This fa- cilitates the mapping of output transliterations from the model to their existing reference transliterations. Experiments demonstrate the superiority of the pro- posed approach in comparison to the state-of-the-art techniques.

As part of future work, we plan to investigate the end-to-end forward and backward transliteration tasks in a unified architecture. We also plan to improve the performance when transliterating words with varying pronunciation to their written forms and are currently working on populating the Elastic Search stack using the verified model predictions.

References

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2017. Arabic machine transliteration using an attention-based encoder-decoder model. *Procedia Computer Science*, 117:287–297.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sukalpa Chanda, Umapada Pal, Katrin Franke, and Fumitaka Kimura. 2010. Script identification–a han and roman script perspective. In *2010 20th international conference on pattern recognition*, pages 2708–2711. IEEE.

Ciprian Chelba and Frederick Jelinek. 2000. Structured language modeling. *Computer Speech & Language*, 14(4):283–332.

Yaping Deng, Lu Wang, Hao Jia, Xiangqian Tong, and Feng Li. 2019. A sequence-to-sequence deep learning architecture based on bidirectional gru for type recognition and time location of combined power quality disturbance. *IEEE Transactions on Industrial Informatics*, 15(8):4481–4493.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.".

Vishal Goyal and Gurpreet Singh Lehal. 2009. Hindi-punjabi machine transliteration system (for machine translation system). *George Ronchi Foundation Journal, Italy*, 64(1):2009.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine trans- lation. In *Proceedings of the AAAI Conference on Ar- tificial Intelligence,* volume 32.

Eric Henke, Ellen M Kaisse, and Richard Wright. 2012. Is the sonority sequencing principle an epiphenomenon. *The sonority controversy*, 18:65–100.

Yuxiang Jia, Danqing Zhu, and Shiwen Yu. 2009. A noisy channel model for grapheme-based machine transliteration. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009),* pages 88–91.

In-Ho Kang and Gil Chang Kim. 2000. English-to-korean transliteration using multiple unbounded overlapping phoneme chunks. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computa- tional linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Soumil Mandal and Karthick Nanmaran. 2018. Normalization of transliterated words in code-mixed data using seq2seq model & levenshtein distance. *arXiv preprint arXiv:1805.08701*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems,* 27.

Rebecca Treiman and Andrea Zukowski. 1990. Toward an understanding of english syllabification. *Journal of Memory and Language*, 29(1):66–85.

Usman Mohy ud Din. 2019. Urdu-english machine transliteration using neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

The Relationship between L2 Grit and Intrinsic Reading Motivation of Filipino Pre-service Teachers in Central Mindanao

Jannie Faye S. Nacario
Department of Professional Education, Notre Dame University, Notre Dame Avenue, Rosary Heights 2, Cotabato City, Philippines
jfayenacario@gmail.com

King Arman A. Calingasan
Life Academy International
CCF Worship and Training Center, Pasig City, Metro Manila, Philippines
kingarmancalingasan@gmail.com

Omar K. Pendi
Department of Professional Education, Notre Dame University, Notre Dame Avenue, Rosary Heights 2, Cotabato City, Philippines
omar.k.pendi@gmail.com

Bai Sittie P. Maguing
Department of Professional Education, Notre Dame University, Notre Dame Avenue, Rosary Heights 2, Cotabato City, Philippines
baisittiemaguing@gmail.com

Abstract

This quantitative study analyzed the levels of and the relationship between the second language (L2) grit and intrinsic reading motivation of pre-service teachers majoring in English language education (N=128) and elementary education (N=108) from two universities in Central Mindanao, Philippines. Using a quantitative correlational research design and a cross-sectional survey method, the randomly selected respondents answered the L2 Grit scale and Intrinsic Reading Motivation Scale which both had good internal consistency. The results from the descriptive statistics showed that both groups had high levels of L2 grit and intrinsic reading motivation. Moreover, these variables had a significant positive correlation based on the Pearson product-moment correlation analyses. This means that when the level of students' grit in learning the second language increases, their motivation to read English texts also increases, and vice-versa. Such also indicates that strengthening the intrinsic reading motivation of the learners will most likely encourage the development of their L2 grit. As a non-cognitive concept, grit assists students in accomplishing the long-term goals they have. Hence, pedagogical implications and recommendations for future study are presented.

1 Introduction

One of the learners' personalities that are extensively studied by educational psychologists is grit. Duckworth et al. (2007) introduced grit to delineate a person's passion and tenacity to pursue a long-term goal amid challenges and adversities. It is a combination of enthusiasm and persistence in the pursuit of a goal that takes a long process before its fulfillment. Duckworth and Gross (2014) argued that grit is one of the valuable determinants of success. Educators believe that studies focusing on grit direct creative processes toward the production of successful students (Keegan, 2017). In the context of the second language (L2) learning and teaching, the second language-domain-specific grit (henceforth L2 grit) has been recently considered an important personality trait of successful L2 learners because its role is relatively new in this realm (Teimouri et al., 2020).

Concerning language learning success, intrinsic motivation may be regarded as correlative to L2 grit. As defined by Ryan and Deci (2000) based on the self-determination theory, intrinsic motivation is an inner zeal to do an activity that elicits internal gratification. This means that intrinsically motivated learners possess an interest and desire to learn something that gives them personal satisfaction, and they are likely to be successful in the academic endeavor (Teimouri et al., 2020). Moreover, Anjomshoa and Sadighi (2015) asserted that successful English language learning requires a learner to be intrinsically motivated. Learners who have high intrinsic motivation seem to have similar characteristics to gritty students. Both manifest an attitude of resilience, perseverance, and sustained interest in learning without expecting external rewards.

To measure students' grit in learning a second language, Teimouri et al. (2020) constructed and validated a second language-domain-specific grit scale. They also examined the relationship between L2 grit and language achievement in a sample of 191 Persian students who studied an English Translation course. They found that gritty students are more passionate about L2 learning and cognitively engaged in class discussion than less gritty students. Moreover, their analyses indicated a positive relationship between L2 grit and language achievement. Similarly, Alamer (2021) validated his newly developed L2 grit scale which he tested among 213 Saudi students who were studying English as an L2. After analyzing the psychometric properties of his self-constructed survey, Alamer (2021) affirmed that the "L2-grit scale is reliable, valid, and suitable for use in L2 research" (p. 1). Freiermuth et al. (2021), on the other hand, interviewed eight gritty English language students from Japan, Malaysia, Taiwan, and Thailand to detail the characteristics of a gritty L2 learner. They discovered that L2 learners who are gritty have the endurance it takes to learn the English language and delight in the learning process. They are never bored by it and are confident in their ability to communicate in it even if they are not yet fluent.

Because L2 grit has been lately explored in L2 learning and teaching, only a few studies on this research topic have provided empirical data. To date, most L2 grit studies have focused on the development and validation of the L2 grit scale (e.g., Alamer, 2021; Teimouri et al., 2020). Consequently, the present study adopts and modifies the available L2 grit scale from the previous research to explore the relationship of this personality trait with a different yet related construct in L2 learning, i.e., intrinsic reading motivation. Another research gap was found in the study of Freiermuth et al. (2021). Although they identified enjoyment or intrinsic motivation as one of the factors that influence L2 grit, they failed to specify what type of intrinsic motivation was referred to by the participants. In addition, Freiermuth et al. (2021) determined motivation as a factor through qualitative analysis only. This psychological construct is chosen because to the best of our knowledge, no efforts have been made to examine the relationship between L2 grit and intrinsic reading motivation. Furthermore, the current research addresses the issue of homogeneity of the sample as one of the limitations pointed out by Alamer (2021). According to him, future researchers must consider using samples from different language learning environments. Therefore, this study is conducted in the Philippines, a multicultural and multilingual learning context where English is considered a second de jure official language. Unlike past studies, this research analyzes the constructs mentioned above using the survey responses from two different groups of language learners (i.e., students majoring in English and students studying elementary education). Diversity of the language learning contexts and language learners may provide an in-depth understanding of grit as an L2 learning construct and the utilization of the L2 grit scale (Alamer, 2021).

1.1 Theoretical Framework

In language learning and academic success, personality and motivation are two of its most important determinants (Kelsen & Liang, 2019). Termed as a trait, Roccas et al., (2002, p. 790) defined personality as "what people are like" and that it may be positive or negative. Specifically, it deals with a collection of underlying traits that determine the actions, thoughts, and feelings of an individual (Medford & McGeown, 2012). Recently, Brandt et al. (2021) have described broadly one's personality as to how one behaves towards something. Concerning the present study, grit and motivation may be commonly mistaken as constructs with an entirely similar identity. Although previous studies have shown a relationship between the said variables, their distinct roles in the completion of goals must be emphasized. First, grit, according to Duckworth et al. (2007), has been considered a personality that entails the capacity to persevere and maintain dedication toward a long-term objective. Second, they suggested the possibility of grit assuming a narrow component of the five-factor model of personality.

The Big Five framework proposes five primary factors of personality that account for the individual differences among people (Medford & McGeown, 2012). In the academic setting, it influences not only the learners' accomplishments but also their language learning as the same framework draws the individual disparity by determining one's attitude, cognition, motivation, temperament, and learning styles (Kelsen & Liang, 2019). The five dimensions of personality include (1) agreeableness, (2) extraversion, (3) neuroticism, (4) openness to experiences, and (5) conscientiousness. Through various contexts, earlier researchers have utilized the same model to predict individual differences (Roccas et al., 2002). Moreover, among the identified personality factors, conscientiousness is most closely associated with grit. Credé et al. (2017) found substantial evidence for the link between grit and conscientiousness in their study. This corroborated the argument of Roberts et al. (2012) that despite being separately developed constructs, the two variables have clear relationships.

According to Roberts et al. (2009), individual differences in terms of one's inclination to be industrious, obedient, organized, responsible, and self-controlled are best described in a spectrum of constructs called conscientiousness. They also postulated that in its hierarchical structure, the upper level can be separated into two aspects: proactive and inhibitive. On the lower level of proactive conscientiousness, however, resides industriousness. Roberts et al. (2012) described industrious people as those who work hard, strive for excellence, and persevere despite obstacles. This description is similar to that of grit, particularly its subcomponent, that is, perseverance of effort. In the study by MacCann et al. (2009), perseverance of effort was found to have a positive relationship with the industrious facet of conscientiousness. In simpler words, grit is among the essential features of human personality and has significant behavioral implications (Costa & McCrae, 1992, as cited in Komarraju & Karau, 2005).

As mentioned earlier, motivation plays a crucial role in learning a target language. Brandt et al. (2021) described motivation as the cause for people's actions towards something. Medford and McGeown (2012) asserted that there exist various theories of motivation, but intrinsic and extrinsic are most often employed in reading studies, which is also the focus of this endeavor. This explains that an intrinsically motivated reader finds such activity to be fundamentally engaging or delightful. On the other hand, an extrinsically motivated reader takes part in a reading activity because of separable factors, for example, receiving a reward or avoiding a penalty. Furthermore, Schiefele et al. (2012) found that high domain-specific intrinsic motivation is equivalent to high success expectations and subsequent desire to demonstrate effort. In addition, intrinsically motivated students are depicted as those who maintain interest as they pursue a personal objective (Ryan & Deci, 2000). This corresponds to a grit subcomponent that is concerned with the consistency of interest (CI).

The relationship between personality and motivation has been evidenced by several studies. In a study by Komarraju and Karau (2005), they argued that high openness to experiences is tantamount to greater academic motivation. Moreover, a similar study yielded results showing conscientiousness and openness to experiences accounting for 17% of the variance in the learners' intrinsic academic motivation (Komarraju et al., 2009). According to other research, an individual's personality may be associated with distinct sub-facets of motivation in various ways. Clark and Schroth (2010) reported that intrinsically motivated students, in terms of acquiring knowledge and completing tasks, were both conscientious and agreeable. On the one hand, learners who were intrinsically motivated toward simulation experiences were likely under the personality factor of openness to experience. Although these past studies have found a link between personality and motivation, their different roles in terms of accomplishing goals must be considered. Motivation explains why one behaves, whereas grit tells how one behaves (Brandt et al., 2021). This, then, resonates with the assumption of the present study that when students are intrinsically motivated in reading, they also have a high level of L2 grit and vice versa.

1.2 Statement of the Problem

In this study, the respondents are pre-service teachers who specialize in English language education and elementary education. Part of the objectives of the study is to determine levels of their L2 grit and intrinsic reading motivation. Most importantly, this present research aims to examine the relationship between intrinsic reading motivation and L2 Grit of the pre-service teachers from two universities in Central Mindanao, Philippines. It specifically attempts to answer the following research questions:

- What is the level of L2 grit and intrinsic reading motivation of students majoring in English and students studying elementary education?
- Is there a significant relationship between L2 grit and intrinsic reading motivation

of students majoring in English and students studying elementary education?

2 Methodology

2.1 Research Design

The present study employed a quantitative survey research method to derive quantitative descriptions in measuring the relationships between variables of the selected sample population (Creswell & Creswell, 2018). Further, the relationship between L2 grit and intrinsic reading motivation was assessed using survey tools, resulting in numerical data that can be evaluated using statistical processes. In line with this, the study followed the cross-sectional survey to draw information on particular phenomena at one period of time (Kelley et al., 2003).

2.2 Research Setting

The current research endeavor was conducted at two universities in Cotabato City, Philippines. Both academic institutions offer education courses such as Bachelor of Elementary Education (BEEd) and Bachelor of Secondary Education (BSEd) - major in English which are essential for this study. It is appropriate to conduct this study in this setting because the target respondents were available in both schools.

2.3 Research Respondents

The respondents for the present study were 236 Filipino college students (N = 191 female; N = 38 male; N = 7 preferred not to reveal their biological sex) from two higher educational institutions in Central Mindanao: a private and a state university. Out of 384 pre-service teachers majoring in English and 149 in elementary education learning English as a second language (L2), 128 (33%) and 108 (72%) from each group responded respectively. In this study, participating students who were specializing in English were heavily exposed to the said language because most of their course works were related to English pedagogy, literature, and linguistics. Meanwhile, participating students who were studying elementary education took English as a minor subject; as a result, they had lesser exposure to English as compared to the former group of respondents. English as a Second Language (ESL) is a term that is used to refer to specialized methods of teaching the English language to individuals whose first language is not English. The respondents' age ranged from 18 to 26 years old with 19 years old being the modal age, and they were mostly freshmen (38.6%).

Moreover, the study followed a multi-stage clustering by beginning with the identification of clusters or groups, followed by the collection of names of individuals belonging to those clusters, and finally extraction of samples from them (Creswell & Creswell, 2017). Ultimately, the study drew from a random sampling technique to provide everyone an equal chance of being selected (Creswell & Creswell, 2017) and prevent biases.

2.4 Research Instruments

L2 Grit Scale. In measuring the L2 grit of college students, this study adopted the validated survey questionnaire of Teimouri et al. (2020). It consists of two components: perseverance of effort (PE) and consistency of interest (CI) in learning a language. The consistency of interest calculates the interest in studying L2, while perseverance of effort assesses the learners' persistence in achieving their goals in L2 learning. This five-point Likert scale from 1 to 5 (not at all like me to very much like me) was acceptable in the present study as it had an internal consistency of 0.794.

Intrinsic Reading Motivation Scale. The present study modified the Reading Motivation Survey created by Guthrie et al. (2009) to measure students' intrinsic reading motivation. There were originally four variables, i.e., intrinsic motivation, avoidance, self-efficacy, and perceived difficulty. However, only seven items under intrinsic reading motivation in their research with the same Likert response format (from 1= never to 4= always) were adopted. After pre-testing this instrument, the intrinsic reading motivation scale was also found acceptable considering its 0.724 internal consistency.

2.5 Data Collection and Analysis Procedures

Considering the ethical concerns and requirements for conducting research with our target participants, we submitted official letters of request to the College of Education deans of one private and one state university in Central Mindanao. An informed consent letter was included in the correspondence, which detailed the aim, procedures for participation in the study, risks, and benefits, as well as the consent form for the participants. A full printed copy of our survey questionnaire was also included in our submission. After the approval, they sent us the official lists of students enrolled in English and Elementary Education courses.

We, then, immediately conducted a pre-testing of instruments twice among the 23 third-year English major students from one of the targeted schools. The first attempt yielded unreliable results, specifically with the avoidance scale under reading motivation. Thus, the same respondents were requested to answer the same questionnaire again. They were also instructed to accomplish the survey with careful consideration and truthfulness. However, the same problem occurred after the second accomplishment and therefore led to the decision of removing the avoidance scale.

After the pre-testing of the instruments, we sent the web-based survey to the respondents via email. In analyzing the relationship between L2 grit and intrinsic reading motivation, the Pearson product-moment correlation using SPSS was utilized. Before that, however, reverse coding was used to accurately analyze the score of the results. The objective of reverse scoring is to recode responses so that a high score corresponds to a low score on the scale. On a 5-point scale, for instance, a four becomes a two, and vice versa.

3 Results and Discussion

3.1 Levels of L2 Grit and Intrinsic Reading Motivation

Table 1 summarizes the findings from the descriptive analysis of ESL students' and English majors' L2 grit and their intrinsic reading motivation. It reveals that both groups have high levels of intrinsic reading motivation and L2 grit, which implies that whether learners are specializing in English or simply studying English, they are most likely to be gritty and intrinsically motivated to read English texts.

Results concerning L2 grit confirm that students majoring in English have remarkable grit in L2 learning (M = 3.58, SD = 1.027). Considering that they receive substantial inputs in the English language as they learn its fundamentals, such a result is barely surprising. It is expected that English majors will most likely be gritty in learning the English language because it is their specialization. They have been exposed to academic coursework related to English pedagogy, linguistics, and literature since their first year at the university. The same expectation was confirmed in the study of Teimouri et al. (2020) in which English-major students, whose future line of work was heavily dependent on their communicative skills in the English language, had high levels of L2 grit. Additionally, it is important to note that these English majors were under the teacher education program that prepares students for an English language teaching career. Hence, if these students aim to obtain an English teaching job after graduation, they will surely set a long-term goal and persevere to achieve it.

Table 1.
L2 Grit and Intrinsic Reading Motivation of Filipino Pre-service Teachers

	Mean	SD	Interpretation
Pre-service English Teachers			
L2 Grit	3.58	1.027	High
Intrinsic Reading Motivation (IRM)	3.20	0.774	High
Pre-service Elementary Teachers			
L2 Grit	3.39	1.061	High
Intrinsic Reading Motivation (IRM)	3.23	0.792	High

Note: 1.00-2.49 = Low IRM; 2.50-4.00 = High IRM
1.00-2.99 = Low L2 Grit; 3.00-5.00 = High L2 Grit

As Duckworth et al. (2007) explained, pursuing English programs is similar to joining a marathon that requires mastery of the target language despite the frustrations its journey necessarily entails. Thus, in order to develop the necessary skills related to English teaching or learning, one must possess grit. For instance, the pre-service teachers from the study of Zawodniak et al. (2021) confessed awareness of their weaknesses in English language learning, but they were firm in addressing and possibly eliminating their identified weaknesses.

Given that literature on L2 grit is still scarce, the presented assumption can also be drawn from the Commission on Higher Education (CHED) Memorandum Order No. 75, s. 2017 under Article IV, Section 6.3.1 which requires pre-service English teachers to use

English, when teaching language and literature, as a glocal language in a multilingual context. This implies that they are provided multiple opportunities to immerse themselves in practicing the target language, whereas these opportunities may encompass a variety of language learning experiences, depending on where the task at hand sits on the spectrum of its difficulty. Hence, as grittier learners of the English language, English majors are expected to keep thriving until they display their desired level of proficiency in the target language.

Similarly, learners majoring in elementary education have high levels of L2 grit (M = 3.39, SD = 1.061). This connotes that the majority of them may also be as passionate and persevering in learning the second language as the English majors are. A possible reason for this is the goal they have in learning it. Duckworth et al. (2007) emphasized that having not only passion and perseverance but also a long-term goal is the quality of a gritty individual. To achieve that goal, a person must demonstrate a strong desire and resilience despite setbacks or difficulties in the learning process. It can be inferred that these pre-service elementary teachers who have taken English subjects since grade school may have long-term objectives of learning the second language. Although they do not specialize in English like the English majors, these learners showed grit in L2 learning.

The results of this study can be supported by the English language attitude of Filipino adults across professions as surveyed by Mahboob and Cruz (2013). They found that majority of their respondents, across ESL communities, preferred English to be taught as a subject in school and be used as a language of instruction, which led them to claim that "English is the language that is perceived to be worthy of investment" (Mahboob & Cruz, 2013, p. 10). Most Filipinos devote their time to learning this language because English still holds a hegemonic position in the Philippines (Mahboob & Cruz, 2013). English continues to be the language in various societal domains in the country such as education and business and is a key to local and international job opportunities. Additionally, this positive attitude towards English can have an impact on one's behavior. For instance, Hein et al. (2020) claimed that a positive attitude serves as a stimulus for one to take action and manifest perseverance, while others are withdrawing in the face of change and setbacks. They also argued that attitudes toward lifelong learning, as well as general learning strategies, were found to predict one component of grit – the persistence of effort (PE). Thus, when students invest in this language, it may mean that they have the goal to acquire it and develop the necessary linguistic skills no matter how tedious the process is. As a result, the pre-service English and elementary teachers in this study may have developed high L2 grit through the years of studying to improve their English language skills that will help them achieve their career goals.

When it comes to reading, intrinsically motivated learners experience genuine pleasure and maintain interest while doing the said activity. They are further described as those who spend their time and effort, especially when developing a thorough knowledge of the texts they read, while also employing appropriate reading strategies (Hebbecker et al., 2019). In this study, the data above indicate that both English majors (M = 3.20, SD = 0.774) and ESL learners (M = 3.23, SD = 0.792) are highly intrinsically motivated readers of English texts. This corroborates the assumption made specifically on English majors having the tantamount intrinsic reading motivation to their L2 grit because their future careers heavily utilize English as a second language.

As mentioned earlier in this paper, interest can also complement one's reading motivation (Alhamdu, 2015). Moreover, such a claim suggests that learners' motivation to read increases when the text piques their interest. In other words, they are most likely to have higher reading motivation when both are taken into consideration. Based on the definition of intrinsic reading motivation provided by Ryan and Deci (2000), it can be surmised that the respondents of this study have the inner zeal to read English reading materials that give them personal satisfaction.

3.2 Relationship between L2 Grit and Intrinsic Reading Motivation

Table 2 presents the correlation data between L2 grit and intrinsic reading motivation of ESL students with the Pearson correlation coefficient as the statistical treatment. The analysis reveals that the L2 grit and intrinsic reading motivation of the said group have a low positive significant relationship (r = .357, p-value < .000). This suggests that when ESL students have a high level of grit in L2 learning, they will most likely have a high level of intrinsic reading motivation as well.

Table 2.
Relationship Between L2 Grit and Intrinsic Reading Motivation of Pre-service Elementary Teachers

Variables	r	p-value
L2 Grit	0.357	0.000
Intrinsic Reading Motivation		

Note: Significant at the .01 level (2-tailed)
N= 108

Similarly, Table 3 shows the correlation data between the same variables of students specializing in English. The same analysis found a low positive significant relationship (r = .371, p-value < .000) between L2 grit and intrinsic reading motivation, indicating that English majors who tend to have high levels of L2 grit may also become highly intrinsically motivated readers.

Table 3.
Relationship Between L2 Grit and Intrinsic Reading Motivation of Pre-service English Teachers

Variables	r	p-value
L2 Grit	0.371	0.000
Intrinsic Reading Motivation		

Note: Significant at the .01 level (2-tailed)
N= 128

At the beginning of a long-term journey to learn a second language, students must reflect on their interest in and intended effort in doing so (Cavilla, 2017). The results presented above are consistent with the research conducted by Changlek and Palanukulwong (2015). In their study, one major statistical finding revealed a significant and positive correlation between intrinsic and extrinsic motivation and grit among high achievers who are learning English as a foreign language. Particularly between intrinsic motivation and perseverance of effort, the same study found a significant and positive but weak relationship. This positive relationship implies that gritty people are more focused on their goals, and such is demonstrated when they get obsessed, as Lehrer (2011) would describe it, with particular activities like reading, in relation to the present study. We can expect, therefore, that intrinsically motivated language learners have the strength to endure when confronted with a difficult task, i.e., reading extremely challenging texts. Gritty students can overcome their fear of failure as they welcome challenges as part of the learning process. They recognize that mastering the target language requires a lot of reading, spanning from simple to complex materials, along with other activities that develop their skills. This reinforces Keegan's (2017) argument that one's personality and motivation are important determinants in language acquisition and educational accomplishments. Existing studies have also discovered that grit is associated with a variety of beneficial outcomes which include academic motivation (Eskreis-Winkler et al., 2014), persistence in accomplishing difficult tasks (Lucas & Nordgren, 2015), and even in delivering outstanding performances such as nationwide spelling competitions (Duckworth et al., 2010).

However, the low correlation between the variables examined in the present study can possibly be explained by the context in which the L2 grit is measured. This indicates that despite both variables heading in the same direction, they do not necessarily have a linear correlation. It should be emphasized that most of the studies conducted on grit came from the Western culture, i.e., the United States, a country whose society is thought to be individualistic (Hofstede, 2001). In order to investigate the potential variances of grit, cultural theories must be considered. The self-construal theory of Markus and Kityama (1991) was used extensively to explain this phenomenon. This theory describes Western individualistic societies as people who view themselves with an autonomous identity, free of their social context (independent self-construal), and capable of pursuing their own goals (Markus & Kitayama, 1991). Contrastingly, the collective societies from the East, or other cultures, see themselves as inherently interdependent components of society (interdependent self-construal). In other words, they share a fundamental connection with one another. As per empirical evidence, past studies among Asian communities revealed that learners usually invest a lot of their time in studying to broaden their academic achievements and maintain a 'face.' This approach to academic success is greatly encouraged in the East (King, 2015) but not from the end of Western students. Taken as a maladaptive approach by the latter group, they do not recommend it as an effective way of gaining learning success (Elliot & Murayama, 2008).

Datu et al. (2016) further added that studies focusing on other cultures, such as the Asian contexts, are still in the marginalized area considering the very little research done. They also pointed out that the situation within the said community may be different, considering their collective values, social conventions, and traditions. Such major differences necessitate the investigation of the applicability of the concept of grit in a collectivist society. Hence, there is a high possibility that the Western individualistic concept of grit might not be appropriate for collective societies. This, then, calls for a modified model of grit that is more culturally applicable in a collectivist society like the Philippines.

Furthermore, grit entails long periods of time by definition. Given that the present study used a cross-sectional survey, this could have also played a factor in such results. Thus, there is a great possibility that measuring grit for a single period of time, by merely answering the questionnaire, and amongst a collective society may be simply not entirely appropriate.

4 Conclusion

Over the past few years, the growing interest in grit does not seem to slow down anytime soon. Various studies have already highlighted its importance across different domains, including the realm of language learning. Previously conducted research studies have proven that grit, as a non-cognitive construct, can help a learner succeed in achieving their identified long-term goals (Duckworth et al., 2007; Duckworth & Gross, 2014). In fact, motivation has been among the various factors that are associated with grit, particularly in language learning (Changlek & Palanukulwong, 2015; Chen et al., 2020; Feng & Papi, 2020). In line with this, the present study discovered that learners of English as an L2 and student teachers specializing in the said language are highly gritty individuals who are also intrinsically motivated learners. Contrary to the presumptions made at the onset of this study, results demonstrated that their differences in terms of their focus of study do not necessarily determine the level of grit they have in L2 learning. However, as Datu et al. (2016) would suggest, L2 Grit scales need to be culturally sensitive, too. Nevertheless, this does not necessarily negate the link determined between the variables. This can, instead, mean that stimulating students' intrinsic reading motivation can also improve their L2 grit. Additionally, the researchers only used pure quantitative methods which might have different possible results if they used different designs in research, such as qualitative research design or mixed method research design in the current study.

5 Recommendations

The relationship between grit and motivation has been established in previous studies. However, linguistic research particularly focusing on language-specific grit seems scant. A mixed-methods study offering a qualitative perspective on the subject and the participants' complex viewpoints may be done as a follow-up. Moreover, several studies across different fields have been stressing the importance of grit, for it almost always guarantees success among those who possess it. In the landscape of academics, there have been empirical data that prove its crucial role in positive academic outcomes (Datu et al., 2016).

The importance of developing L2 grit is highlighted in this study as it positively correlates with intrinsic reading motivation. Thus, the present research recommends identifying teaching strategies that specifically increase the L2 grit of the students to be tested (Alamer, 2021). Conducting grit intervention studies shall help our teachers plan their teaching strategies or practices that promote L2 grit, as well as intrinsic reading motivation, among the learners. This will further shed light on how teachers can provide the help needed by the students in class.

While the literature provides a few strategies that foster L2 grit in students, Duckworth (2013), in her appearance at the TED Conference, suggested that a growth mindset is a good idea for building grit. This means that teachers can use teaching strategies that promote a growth mindset in learning as it also develops the grit of the students (Zhao et al., 2018). For example, English language teachers should praise the reading effort of the students who positively view effort-ability relationships (Calingasan & Plata, 2022), provide process-focused criticism (Dweck, 2008), help students set and achieve a learning goal instead of a performance goal (Dweck & Yeager, 2019), and give them challenging learning tasks (Grant & Dweck, 2003). Duckworth (2013) believed that fostering a growth mindset among individuals is one of the ways to build grit. In fact, gritty individuals are more successful than those who have higher IQs. According to the study by Schwinger et al. (2009) and Wolters (1998), college students have a set of motivational regulation strategies associated with increased effort, academic performance, and persistence. With this, we can infer that a growth mindset would be a great help to boost the motivational regulation strategies of college students in pursuing their long-term goals and it is also a way to increase their grit at the same time.

Moreover, in order to contribute to the growing body of literature on L2 grit, it is first suggested that an L2 Grit Scale, acknowledging the significant differences among collective societies, be developed. Grit may be a personality that anyone can have, but not everyone views it so similarly. Cultural factors may play a crucial part in how one evaluates grit. Hence, for the purpose of gathering as accurate data as possible, a culturally sensitive L2 grit scale must be proposed, examined, and validated for further use.

Lastly, future language researchers may opt to conduct studies that involve ESL in-service teachers. Teimouri et al. (2020) asserted that investigating language teachers' grit and their motivation in teaching is of equal importance. Given that they have first-hand experience in facilitating an ESL classroom, it would be interesting to see just how gritty they are. This trait can be reflected in their teaching practices or in their professional development in general. As someone who always deals with the element of spontaneity common to the teaching profession, it is also of interest to know how they confront difficult situations that arise within the language classroom. Hence, it would be much better to also conduct an interview that shall allow a more in-depth understanding of the variable being examined.

References

Abdullah Alamer. 2021. Grit and language learning: construct validation of L2-Grit scale and its relation to later vocabulary knowledge. Educational Psychology, 41(5), 544–562. https://doi.org/10.1080/01443410.2020.1867076

Ahmar Mahboob and Priscilla Cruz. 2013. English and mother-tongue-based multilingual education: Language attitudes in the Philippines. Asian Journal of English Language Studies, 1, 2-19.

Alhamdu Alhamdu. 2015. Interest and reading motivation. Psikis: Jurnal Psikologi Islami, 1(1), 1-10. https://doi.org/10.19109/psikis.v1i1.552

Andrew J. Elliot and Kou Murayama. 2008. On the measurement of achievement goals: Critique, illustration, and application. Journal of Educational Psychology, 100, 613–628.

Angela Lee Duckworth. 2013, May. Grit: The power of passion and perseverance. [Video]. TED Conferences.https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseveranc

Angela Lee Duckworth, Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. Grit: perseverance and passion for long-term goals. Journal of personality and social psychology, 92(6), 1087.

Angela Lee Duckworth and James J. Gross. 2014. Self-Control and Grit. Current Directions in Psychological Science, 23(5), 319–325. https://doi.org/10.1177/0963721414541462

Angela Lee Duckworth, Teri A. Kirby, Eli Tsukayama, Heather Berstein, and K. Anders Ericsson. 2010. Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. Social Psychological and Personality Science, 2(2), 174–181. https://doi.org/10.1177/1948550610385872

Ansari Changlek and Thanyapa Palanukulwong. 2015. Motivation and grit: Predictors of language learning achievement. Veridian E-Journal, Silpakorn University (Humanities, Social Sciences and arts), 8(4), 23-36.

Brent A. Kelsen and Hsin-Yi Liang. 2019. Role of the Big Five personality traits and motivation in predicting performance in collaborative presentations. Psychological Reports, 122(5), 1907-1924.

Brent W. Roberts, Joshua J. Jackson, Jennifer V. Fayard, Grant Edmonds, and Jenna Meints. 2009. Conscientiousness.

Brent W. Roberts, M. Brent Donnellan, and Patrick L. Hill. 2012. Personality trait development in adulthood: Findings and implications. In H. Tennen & J. Suls (Eds.), Handbook of psychology (pp. 183–196). New York: Wiley Publishing, https://doi.org/10.1002/9781118133880.hop205009

Brian J. Lucas and Loran F. Nordgren. 2015. People underestimate the value of persistence for creative performance. Journal of Personality and Social Psychology, 109(2), 232.

Carol S. Dweck. 2008. Mindsets: How praise is harming youth and what can be done about it. School Library Media Activities, 24, 55-58.

Carol S. Dweck and David S. Yeager. 2019. Mindsets: A view from two eras. Perspectives on Psychological Science, 14(3), 1-16. https://doi.org/10.1177/1745691618804166

Carolyn MacCann, Angela Lee Duckworth, and Richard D. Roberts. 2009. Empirical identification of the major facets of conscientiousness. Learning and individual differences, 19(4), 451-458.

Christopher A. Wolters. 1998. Self-regulated learning and college students' regulation of motivation. Journal of educational psychology, 90(2), 224.

Commission on Higher Education . 2017. CMO No. 75 s. 2017. https://ched.gov.ph/wp-content/uploads/2017/11/CMO-No.-75-s.-2017.pdf

Derek Cavilla. 2017. The effects of student reflection on academic performance and motivation. Sage Open, 7(3), 2158244017733790.

Emma Medford and Sarah P. McGeown. 2012. The influence of personality characteristics on children's intrinsic reading motivation. Learning and Individual Differences, 22(6), 786-791.

Geert Hofstede. 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage publications.

Hazel R. Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. Psychological review, 98(2), 224.

Heidi Grant and Carol S. Dweck. 2003. Clarifying achievement goals and their impact. Journal of Personality and Social Psychology, 85(3), 541- 553. https://doi.org/10.1037/0022-3514.85.3.541

Jesus Alfonso Daep Datu, Jana Patricia Millonado Valdez, and Ronnel Bornasal King. 2016. The successful life of gritty students: Grit leads to optimal educational and well-being outcomes in a collectivist context. In The psychology of Asian learners (pp. 503-516). Springer, Singapore.

Joanna Zawodniak, Miroslaw Pawlak, and Mariusz Kruk. 2021. The Role of Grit Among Polish EFL Majors: A Comparative Study of 1st-, 2nd-, and 3rd-Year University Students. Journal for the Psychology of Language Learning, 3(2), 118-132.

John T. Guthrie, Cassandra S. Coddington, and Allan Wigfield. 2009. Profiles of reading motivation among African American and Caucasian students. Journal of Literacy Research, 41(3), 317-353.

John W. Creswell and John David Creswell. 2017. Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.

John W. Creswell and John David Creswell. 2018. Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE.

Jonah Lehrer. 2011. Which traits predict success? The importance of grit. Wired.

Karin Hebbecker, Natalie Förster, and Elmar Souvignier. 2019. Reciprocal effects between reading achievement and intrinsic and extrinsic reading motivation. Scientific Studies of Reading, 23(5), 419-436.

Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. Good practice in the conduct and reporting of survey research. International Journal for Quality in health care, 15(3), 261-266.

Kelly Keegan. 2017 Identifying and building grit in language learners. In English Teaching Forum, 55(3), 2-9.

King Arman Calingasan and Sterling Plata. 2022. Effects of effort praise on struggling Filipino ESL readers' mindset and motivation. Indonesian Journal of Applied Linguistics, 11(3). 601-611. https://ejournal.upi.edu/index.php/IJAL/article/view/32898

Lauren Eskreis-Winkler, Angela Lee Duckworth, Elizabeth P. Shulman, and Scott A. Beal. 2014. The grit effect: Predicting retention in the military, the workplace, school and marriage. Frontiers in psychology, 5, 36.

Leila Anjomshoa and Firooz Sadighi. 2015. The importance of motivation in second language acquisition. International Journal on Studies in English Language and Literature (IJSEL), 3(2), 126-137.

Liying Feng and Mostafa Papi. 2020. Persistence in language learning: The role of grit and future self-guides. Learning and Individual Differences, 81, 101904. https://doi.org/10.1016/j.lindif.2020.101904

Malte Schwinger, Ricarda Steinmayr, and Birgit Spinath. 2009. How do motivational regulation strategies affect achievement: Mediated by effort management and moderated by intelligence. Learning and individual differences, 19(4), 621-627.

Marcus Credé, Michael C. Tynan, and Peter D. Harms. 2017. Much ado about grit: A meta-analytic synthesis of the grit literature. Journal of Personality and social Psychology, 113(3), 492.

Mari H. Clark and Christopher A. Schroth. 2010. Examining relationships between academic motivation and personality among college students. Learning and individual differences, 20(1), 19-24.

Mark Freiermuth, Chomraj Patanasorn, Latha Ravindran, and Hsin-chou Huang. 2021. Getting to the Nitty-Gritty of Grit: A Descriptive Characterization of Gritty L2 Learners from Thailand, Malaysia, Taiwan, and Japan. Journal for the Psychology of Language Learning, 3(2), 133–155. https://doi.org/10.52598/jpll/3/2/9

Meera Komarraju and Steven J. Karau. 2005. The relationship between the big five personality traits and academic motivation. Personality and individual differences, 39(3), 557-567.

Meera Komarraju, Steven J. Karau, and Ronald R. Schmeck. 2009. Role of the Big Five personality traits in predicting college students' academic motivation and achievement. Learning and individual differences, 19(1), 47-52.

Naemi D. Brandt, Anne Israel, Michael Becker, and Jenny Wagner. 2021. The joint power of personality and motivation dynamics for occupational success: Bridging two largely separated fields. European Journal of Personality, 35(4), 480-509.

Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary Educational Psychology, 25(1), 54–67. https://di.org/10.1006/ceps.1999.1020

Ronnel B. King. 2015. Examining the dimensional structure and nomological network of achievement goals in the Philippines. Journal of Adolescence, 44, 214–218.

Sonia Roccas, Lilac Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. Personality and social psychology bulletin, 28(6), 789-801.

Ulrich Schiefele, Ellen Schaffner, Jens Möller, and Allan Wigfield, A. 2012. Dimensions of reading motivation and their relation to reading behavior and competence. Reading research quarterly, 47(4), 427-463.

Vello Hein, Andre Koka, Hanna Kalajas-Tilga, Henri Tilga, and Lennart Raudsepp. 2020. The effect of grit on leisure time physical activity. An application of the theory of planned behaviour. Balt J Health Phys Act, 12(1), 78-85. https://doi.org/10.29359/BJHPA.12.1.08

Xinjie Chen, Julie Lake, and Amado M. Padilla. 2020. Grit and motivation for learning English among Japanese university students. System, 96, 102411.

Yasser Teimouri, Luke Plonsky and Farhad Tabandeh. 2020. L2 grit: Passion and perseverance for second-language learning. Language Teaching Research, 136216882092189. https://doi.org/10.1177/1362168820921895

Yukun Zhao, Gengfeng Niu, Hanchao Hou, Guang Zeng, Liying Xu, Kaiping Peng, and Feng Yu. 2018. From growth mindset to grit in Chinese schools: The mediating roles of learning motivations. Frontier Psychology. https://doi.org/10.3389/fpsyg.2018.02007

Pronunciation of English Words with /th/ Sounds among Senior High School Learners

Jeremiah L. Saavedra
Enerdino C. Coronel – Baluno National High School, Baluno, Zamboanga City,
Zamboanga del Sur, Philippines
jeremiahsaavedra88@gmail.com

Abstract

ESL learners find pronunciation one of the most challenging aspects in speaking an English language especially if the phonemes of the target language are not present in their mother tongue sound inventory. This study aimed to determine the variation of the pronunciation of English words with dental fricatives or the /th/ sounds among the Chavacano, Tausug, and Visayan senior high school learners. Cross-sectional design and total enumeration was employed in the data collection, and percentage frequency distribution was used for data analysis. Majority of the 32 learner-participants were rated excellent in pronouncing the words with /θ/ sound but poor in pronouncing the words with /ð/ sound. The Tausug and Visayan speakers tend to pronounce the words with voiceless /th/ sound more accurately than the Chavacano speakers, but the latter was better than the other ethnolinguistic counterparts in pronouncing the words with voiced /th/ sound. Mispronunciation of words made by the learner-participants was an affirmation of Richard's (1974) theory of interlanguage error that signals the negative transfer of phonological elements of the mother tongue into the L2. It can be concluded that the learner-participants needed to practice more in pronouncing the words with fricative consonants to make themselves understood and to avoid miscommunication.

1 Introduction

The world's population is now dominated by approximately 1.5 billion speakers of the English language, either as their lingua franca or as their second language (Lyons, 2021; Szmigiera, 2022). Among the macroskills of language (along with reading, writing, and listening), speaking seems to attract a lot of English learners' attention. The foundation of speaking constitutes phonetics both in theory and practice that results to good pronunciation, making the communication process more intelligible. Thus, accuracy of pronunciation which is claimed to be the most important element of learning oral skills in a second language should be highly regarded (Pennington, 2019). Nowadays, L2 pronunciation has abruptly evolved as an interdisciplinary field unlike several decades ago when language research and pedagogy related to pronunciation seemed to be insignificant, hence neglected (Derwing & Munro, 2010; Nagle et al., 2019; Tergujeff, 2012). The emergence of research on L2 pronunciation has shed light on issues, theories, and possible solutions on how to effectively improve the pronunciation of learners. In fact, many second language (L2) learners struggle in this aspect especially in making themselves understood or understand others.

Haugen (1972, p. 325) defined language ecology as "the study of relationships between a language and the environment, basically identified by the speakers who learn and use it, and transmit to others". Creese and Martin (2003) supported this ideology by adding that investigation is required to determine the interrelationship of languages in a specific society, the speakers of the language, and the social structures in which the languages are spoken in a society. This means that language learners who have a high exposure to a target language are more likely to have an accurate pronunciation or oral skill than learners who have less exposure.

Some English consonants cause pronunciation problems to L2 speakers. Pronouncing a group of two successive letters that produces a single sound called "digraph" especially in /th/ phonemes is considered one of their difficulties, i.e., the voiced dental fricative /ð/ (as in *this*) and the voiceless dental fricative /θ/ (as in *thing*). Since the two constrastive phonemes are not present in the Filipino language or in any local language of the Philippines, most Filipino ESL speakers tend to mispronounce the English words, eventually changing the meaning of the words completely, e.g., pronouncing "day" /deI/ instead of "they" /ðeI/.

One thing to consider is the absence of digraphs in the lingua francua sound inventory of the learners which, in contrast, is present and often used in their target language. In fact, Yamaguchi (2014) claims that /th/ sounds are not found in the sound inventory of any local language in Malaysia, that is why L2 learners often mispronounce the words with fricative consonant in the target language. Moreover, Adila and Refnaldi (2019) claimed that there were six kinds of consonant errors in the senior high school students' speaking performance when they conducted a study in an Indonesian university. One factor was because the pronunciation was generally influenced by the speakers' mother tongue. Lastly, ineffectiveness of the phonetic teaching process especially in the primary years of language learning is identified to be another reason of pronunciation problem (Almutalabi, 2018). All these claims with regard to pronunciation errors are possibly anchored from the two types of language errors (Richards, 1974). The interlingual error which is primarily caused by the interference of the mother tongue in learning the target language, and the intralingual error which is caused by transferring the L1 language rules into the target language. Keshavarz (1999) claims that interlanguage errors occur at various levels such as the transfer of linguistic elements of the lingua franca into the L2. Meanwhile, intralingual errors happen when language learners over-generalize, ignore the restriction of language rules, incompletely apply the rules, and assume the concepts wrongly.

Gilakjani (2011) posits the following reasons why L2 learners commit mistakes in pronouncing words in their target language. First, learners have and show no interest. Second, learners have less exposure in their target language. Lastly, language teachers do not emphasize the importance of teaching pronunciation nor they have the sufficient materials in teaching pronunciation effectively. Apparently, it is important for language teachers to draw attention to the importance of teaching pronunciation correctly and apply the contextualized strategies in the language classroom to motivate the language learners in practicing their oral skills accurately.

Enerdino C. Coronel – Baluno National High School is a small secondary school located in the rural hills of the western part of Zamboanga City, Philippines. From 2016 to 2022, its total student population across junior high school and senior high school levels has gradually increased from 250 to 370 learners. The Senior High School Department has an average of 30 to 45 learners per level. Majority of them come from the ethnolinguistic groups of Chavacano and Tausug, and only few learners speak the Cebuano/Visayan language.

In their language subjects (English and Filipino), it was observed that most of the learners mispronounce the words and letters presented and used in their target language. This gave an interest to the researcher to investigate the senior high school learners' pronunciation of words with /ð/ and /θ/ sounds in the English language according to their ethnolinguistic groups.

1.1 Statement of the Problem

This study sought to find the variation on the pronunciation of English words with /th/ sounds particularly the /**ð**/ (voiced dental fricative) and the /**θ**/ (voiceless dental fricative) among the Grade 11 senior high school learners of Enerdino C. Coronel - Baluno National High School. Specifically, the study aimed to answer the following questions:

1. What was the overall result of the pronunciation of English words with dental fricatives or the /th/ sounds?
2. What were the variations on the pronunciation of English words with voiced and voiceless /th/ sounds among the senior high school learners in terms of their ethnolinguistic groups?

1.2 Significance of the Study

This study will benefit the ESL instructors especially the language teachers assigned in rural schools. Public schools in far-flung areas usually house diverse ethnolinguistic groups of learners and have less exposure to their L2. Language teachers need to give extra attention to teaching correct pronunciation to their learners during actual classes, interventions, or anytime when L2 is used as this will impact their oral communication process inside the classroom and in other context.

This study will benefit the ESL learners to be aware of the possible problems with regard to their pronunciation of English words with phonemes that are not present in their lingua franca. Language learners should know that pronunciation is an essential aspect of oral skill that gives lasting impression to their listeners.

Lastly, this study will benefit the Education Program Supervisors for the language subjects under the Department of Education (DepEd), the language practitioners, and other related positions in the education sector. The results of this study will help them suggest appropriate interventions, activities, and programs that will enable and hone the pronunciation accuracy of the language learners regardless of their ethnolinguistic background.

2 Methods

2.1 Design

The researcher used cross-sectional design as the primary aim was to collect the data from the senior high school learners at a specific point in time. Thomas (2020) states that this type of research design is appropriate if the purpose was to observe the chosen participants without influencing them. There is no need for follow-ups or interventions after the data has been collected, thus cost-effective. Nevertheless, cross-sectional design can only employ a relatively passive approach to making causal inferences based on the actual findings.

Total population or complete enumeration was the technique for data collection. It is a type of purposive sampling technique in which the researcher chose to examine the entire population that has a particular set of characteristics such as specific attributes or traits, experience, knowledge, skills, exposure to an event, etc. In this research, the characteristics referred to the ethnolinguistic groups of the learner-participants which influenced the researcher’s choice of total population sampling based on the following reasons:

a. **The population size was relatively small**. Unlike the other public senior high schools that usually have more than one section per grade level and offered more than one strand, Enerdino C. Coronel – Baluno National High School which is a small rural school only opens one section per grade level in the upper secondary level. Because of its small population, ECC-BNHS only offers the Humanities and Social Sciences (HUMSS) strand ever since the K to 12 curriculum of Department of Education was implemented in the Philippines.

b. **The population had diverse ethnolinguistic groups**. Despite the small population of the school, there were three ethnolinguistic groups among the senior high school learners and these were Chavacano, Tausug, and Bisaya.

2.2 Respondents

The respondents for this study were the Grade 11 senior high school learners of Enerdino C. Coronel – Baluno National High School (ECC-BNHS), Baluno, Zamboanga City. They were enrolled in Humanities and Social Sciences strand. Their class was composed of 32 learners; 14 males and 18 females.

Among the 14 males, five of them spoke Chavacano, six spoke Tausug, and three spoke the Visayan language. On the other hand, 18 females were comprised of eight Chavacano speakers; seven Tausug speakers, and three Visayan speakers. Overall, there were thirteen (13) Chavacano speakers, thirteen (13) Tausug speakers, and six (6) Visayan speakers.

2.3 Instruments

Three research instruments were utilized in this study: a) a modified form that contained the information of the senior high school learner-participants; b) a pronunciation practice sheet that contained words with voiced and voiceless fricative consonants; and c) the actual pronunciation test tool to determine the pronunciation accuracy of the senior high school learners on English words with both voiced and unvoiced /th/ sounds. The language teacher used a scoring rubric to evaluate the accuracy of the /th/ sound pronunciation of the learner-participants, and a mobile phone audio recorder to recheck the pronounced words with /th/ sounds before the data were analyzed.

Modified School Form 1

The modified form was designed to collect the personal information of the senior high school learner-participants, adapted from the School Form 1 (SF1) also known as the School Register. Unlike the original School Form 1 that collected information such as their complete name; sex; birth date; age (as of first Friday of June in the school year enrolled); mother tongue; ethnic group; religion; home address; parents' names; guardian's name and relationship; and the contact number of the parent or guardian, the form modified by the researcher collected only the learner-participants' necessary information relevant to the study such as their age, sex, home address, language spoken at home (L1) and other languages they speak.

Pronunciation Practice Sheet

This instrument contained English words with voiced and voiceless dental fricatives adapted from an American English-based textbook titled *Phoenix Language Package: Skill Builders for English Proficiency* for secondary level. It was stated in the first paragraph of the textbook's Foreword that the textbook aimed to improve the listening, speaking, and writing skills (including grammar) of its readers.

The pronunciation practice tool used by the learner-participants to practice the English words with voiceless and voiced /th/ sounds was divided into three columns: 10 words with initial and medial voiceless /th/ sound in the first column, 10 words with voiceless and voiced /th/ sound in the second column, and 10 words with medial and final voiced /th/ sound in the third column. Overall, there were 30 English words with fricative consonants in the pronunciation practice sheet.

Actual Pronunciation Test Tool

This instrument was a minimal-pair reading aloud test on both voiced and voiceless /th/ sounds taken from the same textbook titled *Phoenix Language Package: Skill Builders for English Proficiency*. Since the textbook contained the pronunciation of English words in minimal pairs with dental fricative sounds, the researcher proposed that the selected material was relevant to achieve the objective of this study. Hence, the third instrument which was based from the textbook with American English content was a valid research tool to measure the overall result of the senior high school learner-participants' pronunciation of words with dental fricatives or the /th/ sounds used in the English language.

The actual pronunciation test tool was split into two sets: Set A that comprised 10 English words with voiceless dental fricative /θ/ and 10 words with /t/ sound; and Set B that comprised 10 English words with voiced dental fricative /ð/ and 10 words with /d/ sound. Overall, there were 40 English words in the pronunciation test tool, but only the list of words under /θ/ column and /ð/ column were evaluated by the language teacher since that was the primary scope of the study.

SET A Words with /θ/ and /t/		SET B Words with /ð/ and /d/	
/θ/	/t/	/ð/	/d/
think	tink	than	Dan
thank	tank	then	den
through	true	lather	ladder
three	tree	thine	dine
thick	tick	thy	die
thrust	trust	though	dough
thyme	time	those	dose
thought	taught	they	day
thread	tread	these	dizzy

Table 2: Instrument used to determine the pronunciation errors of the learner-participants in both voiced and voiceless /th/ sounds

Rubric and Audio Recorder

Like any other evaluation tool, rubrics are useful to assess performances. In this case, a rubric is used to determine the pronunciation accuracy of the senior high school learner-participants specifically on the sets of English words with voiced and voiceless /th/ sounds. The resulting judgment of quality based on this tool contains within it a description of the actual performance that can be used for feedback. The rubric used in this study has two major aspects: coherent sets of criteria and descriptions of levels of performance for these criteria. The criteria and performance-level descriptions help the senior high school learner-participants understand their actual pronunciation of the English words with /th/ sounds and how these words were supposed to pronounce.

To assess the oral reading pronunciation of the learner-participants, holistic rubric was used. It consisted of a single scale with all criteria to be included in the evaluation being considered together.

When a senior high school learner-participant pronounced all the words with dental fricative correctly in each set, or made minimal errors with at least one to three mispronounced English words with /th/ sounds out of 10 per set, he or she would be rated "Excellent". When he or she mispronounced the English words four to six times out of 10 per set, he or she would be rated "Average". Lastly, when he or she mispronounced the English words with /th/ sounds seven to nine times per set, or mispronounced all words with /th/ sounds per set, he or she would be rated "Poor".

To recheck the pronunciation of words made by the senior high school learner-participants, the English language teacher recorded their utterances via Oppo A37 built-in audio recorder during the conduct of the data gathering process. The device would identify the possible pronunciation errors on the dental fricative sounds whenever the language teacher misheard the utterance of the learner-participant during the actual pronunciation test.

2.4 Data Gathering Procedure

Prior to the conduct of the study, the researcher sent a letter of permission to the school principal. The data collection was conducted during the first period class of the Grade 11 senior high school learners. It took one hour and thirty minutes to accommodate the 32 senior high school learner-participants who were all present during the data gathering schedule.

At the beginning of the data collection, the senior high school learner-participants were asked to accomplish the SF1 modified form. After the form was completely accomplished, each learner-participant was given a pronunciation practice sheet then they read aloud the English words with voiceless and voiced /th/ sounds together with their English language teacher.

After the pronunciation drill, the learner-participants were given a copy of the actual pronunciation test tool. The language teacher informed them that only the English words with both voiced and voiceless dental fricative sounds would be evaluated, disregarding the words with /t/ sound in Set A and the words with /d/ sound in Set B. Individually, they proceeded to the actual pronunciation test on minimal pairs of English words with voiced and voiceless dental fricative sounds along with the English words with /t/ and /d/ sounds. The process was facilitated and evaluated by the language teacher who was an English major.

The actual oral reading test was audio-recorded to identify the misheard pronunciation of the senior high school learner-participants. The language teacher required them to read orally the English words with voiced /th/ and /t/ sounds twice in set A and read the English words with voiceless /th/ and /d/ sounds twice in set B.

The language teacher ticked on the empty spaces before the English words found on the actual pronunciation test sheet when the learner-participants pronounced the words with /th/ sounds correctly. The spaces were left unmarked when the English words with /th/ sounds were mispronounced.

After the data collection, each senior high school learner-participant's audio recording was carefully listened by the language teacher to recheck the pronunciation of the English words with dental fricatives. No bearing was given to any mispronounced words with /t/ sound in Set A and /d/ sound in Set B as these two sounds were not the focus of the study. The pronunciation test results in both voiced and voiceless /th/ sounds were then statistically treated according to the ethnolinguistic groups of the senior high school learner-participants.

2.5 Data Analysis

The researcher employed percentage frequency distribution on the data gathered to further analyze the results of the pronunciation accuracy of the senior high school learner-participants of Enerdino C. Coronel – Baluno National High School. Percentage frequency distribution is defined as a display of data that signifies the percentage of observations that exist for each data point or grouping of data points (Lavrakas, 2008). This method is useful to express the relative frequency of the accurately pronounced English words with voiced and voiceless dental fricatives. After the results on the pronunciation of English words on voiced and voiceless dental fricatives were analyzed, the researcher interpreted the data based on the adjectival descriptions of the rubric for the actual pronunciation test.

3 Results

To determine the results of the study, the actual oral reading test in two sets of minimal pairs was the basis for assessing the pronunciation accuracy on the English words with voiced dental fricative and unvoiced dental fricative uttered by the senior high school learner-participants.

Among the five male Chavacano speakers, three (3) were rated excellent while two (2) were poor in pronunciation of English words with voiceless dental fricative /θ/ sound in minimal pairs. No male Chavacano speaker was rated average in the given oral reading test. Meanwhile, among the six male Tausug speakers, three of them were rated excellent and the other three were rated poor in the pronunciation of the words with voiceless /th/ sound. Like the male Chavacano speakers, no male Tausug speaker was rated average. Lastly, among the male Visayan speakers, one (1) of them was rated excellent, the other one was average, and the last one was poor in the pronunciation test.

On the other hand, among the eight (8) female Chavacano speakers, four (4) were rated excellent, 1 was average, and 3 were poor in the pronunciation of the English words with voiceless dental fricative sound in Set A minimal pair. Among the seven (7) female Tausug speakers, five (5) were rated excellent and two (2) were poor in the reading aloud test. No female Tausug speaker had an average rating in the pronunciation of the English words with unvoiced dental fricative sound. Lastly, all of the three Visayan speakers were rated excellent in the minimal pair assessment.

To answer the research question 1 which seeks to determine the overall result of the pronunciation of the English words with voiceless fricative consonant, 19 out of 32 senior high school learner-participants were rated excellent in pronouncing the unvoiced dental fricative or the /θ/ sound. This means that more than half of the total sample size pronounced the 10 English words with unvoiced dental fricative sound correctly, or at least committed minimal errors with at least one to three mispronounced English words in unvoiced /th/ sounds. Meanwhile, only two (2) out of 32 senior high school learner-participants were average in the pronunciation of 10 English words with unvoiced /th/ sounds. These two speakers (regardless of their ethnolinguistic background) committed at least four to six pronunciation errors in the English words with unvoiced /th/ sounds in Set A minimal pairs. Lastly, 11 out 32 senior high school learner-participants were poor in pronouncing the 10 English words with unvoiced /th/ sounds. These speakers mispronounced 7 to 9 English words with voiceless /th/ sounds in Set A minimal pairs, or possibly pronounced almost all the English words with unvoiced /th/ sounds incorrectly.

SHS Learners	Excellent /θ/ sound	Average /θ/ sound	Poor /θ/ sound	Total
MALE				
Chavacano	3	0	2	5
Tausug	3	0	3	6
Visayan	1	1	1	3
FEMALE				
Chavacano	4	1	3	8
Tausug	5	0	2	7
Visayan	3	0	0	3
TOTAL	19	2	11	32

Table 3: Results of the senior high school learners' assessment on pronunciation of the English words with voiceless /th/ sound

To answer the research question number 2 that seeks to determine the variation of the pronunciation of the English words with unvoiced fricative consonant or the /θ/ sound according to the ethnolinguistic groups, 21.9% of the total sample size were the Chavacano senior high school learner-participants who were rated excellent in pronouncing the 10 English words with unvoiced /th/ sounds. This means that more than half of the Chavacano speakers pronounced all the English words with voiceless dental fricative sound correctly, or committed minimal errors with one to three mispronounced words in Set A minimal pairs.

Moreover, 25.0% of the total sample size were the Tausug senior high school learner-participants who were excellent in the pronunciation of the English words with unvoiced dental fricative sounds. This implies that most Tausug speakers were likely to have accurate pronunciation of English words with unvoiced /th/ sounds.

Finally, 12.6% of the total sample size or four (4) out of six (6) Visayan senior high school learner-participants were excellent in the pronunciation of English words with unvoiced /th/ sounds. This also signifies that almost all of the Visayan speakers were more likely to pronounce all the English words with unvoiced /th/ sounds correctly rather than committing pronunciation errors.

Overall, 59.5% of the Grade 11 senior high school learner-participants were excellent, 6.2% were average, and 34.3% were poor in the pronunciation of the English words with voiceless dental fricative sound. This means that more than half of the senior high school learner-participants pronounced the English words with /θ/ sound with at least one to three errors, or possibly pronounced all the 10 English words with /θ/ sound correctly. Also, more than half of each ethnolinguistic group only committed minimal errors in the pronunciation of the English words with /θ/ sound or never mispronounced the 10 English words with /θ/ sound at all. Conversely, more than a third of the senior high school learner-participants needed to improve their pronunciation of English words with /θ/ sound as they mispronounced the words seven to nine times in Set A minimal pairs, or possibly pronounced all the 10 English words with voiceless /th/ sounds incorrectly.

SHS Learners	Excellent /θ/ sound	Average /θ/ sound	Poor /θ/ sound	Total
Chavacano	7 (21.9%)	1 (3.1%)	5 (15.6%)	13 (40.6%)
Tausug	8 (25.0%)	0	5 (15.6%)	13 (40.6%)
Visayan	4 (12.6%)	1 (3.1%)	1 (3.1%)	6 (18.8%)
Total	19 (59.5%)	2 (6.2%)	11 (34.3%)	32 (100%)

Table 4: Consolidated results of the senior high school learners' assessment on pronunciation of the English words with voiceless /th/ sound

For the pronunciation of the English words with voiced /th/ sound, among the five (5) male Chavacano speakers, two (2) of them were rated excellent; one (1) was rated average; and the other 2 were rated poor. Three (3) were average and 3 were poor in the oral reading test among the six male Tausug speakers. No male Tausug speaker was rated excellent in pronouncing the English words with voiced /th/ sound. Lastly, among the 3 male Visayan speakers, each of them were excellent, average, and poor respectively.

On the other hand, among the eight (8) female Chavacano speakers, four (4) were rated excellent, one was average, and 3 were poor in pronouncing the English words with voiced dental fricative sound. 2 were excellent, 3 were average, and 2 were poor among the seven (7) female Tausug speakers in their pronunciation of English words with voiced dental fricative sounds. Lastly, among the 3 female Visayan speakers, 1 was excellent and the remaining 2 were poor in the oral reading test. No female Visayan speaker was rated average in the Set B minimal pair reading aloud assessment.

To answer the research question number 1 that seeks to determine the overall result of the pronunciation of the English words with voiced fricative consonant, ten (10) out of 32 senior high school learner-participants were rated excellent, nine (9) were average, and thirteen (13) were poor in the pronunciation of English words in Set B minimal pairs. This implies that more than a third of the total sample size need to improve their pronunciation of the English words with the voiced /th/ sounds in Set B minimal pairs. Apparently, they mispronounced the set of English words with voiced /th/ sound seven to nine times, or mispronounced almost all the English words with voiced dental fricative sounds.

SHS Learners	Excellent /ð/ sound	Average /ð/ sound	Poor /ð/ sound	Total
MALE				
Chavacano	2	1	2	5
Tausug	0	3	3	6
Visayan	1	1	1	3
FEMALE				
Chavacano	4	1	3	8
Tausug	2	3	2	7
Visayan	1	0	2	3
TOTAL	10	9	13	32

Table 5: Results of the senior high school learners' assessment on pronunciation of the English words with voiced /th/ sound

To answer the research question number 2 that seeks to determine the pronunciation of the English words with voiced fricative consonant according to the ethnolinguistic groups, 18.8% (six speakers) were rated excellent, 6.25% (two speakers) were average, and 15.6% (five speakers) were poor among the 13 Chavacano senior high school learner-participants in pronouncing the English words with voiced /th/ sound. This implies that almost half of the male and female Chavacano senior high school learner-participants pronounced the English words with voiced dental fricative sounds with at least one to three pronunciation errors, or almost half of the Chavacano speakers pronounced the English words with voiced /th/ sounds almost perfectly. Meanwhile, five speakers out of 13 male and female Chavacano senior high school learner-participants committed seven to nine pronunciation errors in the 10 English words with voiced dental fricative sounds, or five out of 13 Chavacano speakers mispronounced almost all the 10 English words with voiced /th/ sounds in Set B minimal pairs.

Among the 13 Tausug senior high school learner-participants, 6.25% (two speakers) were excellent, 18.8% (six speakers) were average, and 15.6% (five speakers) were poor in the oral reading test. This means that only two out of 13 Tausug speakers pronounced the 10 English words with voiced /th/ sounds almost perfectly, or only two out of 13 Tausug speakers committed minimal errors in the pronunciation of 10 English words with voiced /th/ sounds in Set B minimal pairs. Meanwhile, almost half of the male and female Tausug speakers committed at least four to six errors in the pronunciation, thus rated average. Five out of 13 Tausug speakers committed seven to nine pronunciation errors on the 10 English words with voiced /th/ sounds, or five out of 13 Tausug speakers mispronounced almost all the 10 English words with voiced /th/ sounds in Set B minimal pairs.

Finally, among the six Visayan senior high school learner-participants, 6.25% (two speakers) were excellent, 3.12% (one speaker) was average, and 9.37% (three speakers) were poor in the pronunciation of English words with voiced dental fricative sounds in Set B minimal pairs. This implies that half of the Visayan senior high school learner-participants needed to improve their pronunciation of English words with voiced dental fricatives as they committed seven to nine pronunciation errors on the 10 English words in Set B minimal pairs, or half of the Visayan speakers mispronounced almost all the 10 English words with voiced dental fricative sounds.

Overall, 40.6% of the Grade 11 senior high school learner-participants need to improve their pronunciation on the English words with voiced /th/ sounds in Set B minimal pairs. More than a third of the total sample size mispronounced almost all the 10 English words with voiced dental fricative sounds, committed at least seven to nine pronunciation errors, or possibly no correct pronunciation at all.

SHS Learners	Excellent /ð/ sound	Average /ð/ sound	Poor /ð/ sound	Total
Chavacano	6	2	5	13
	(18.8%)	(6.25%)	(15.6%)	(40.6%)
Tausug	2	6	5	13
	(6.25%)	(18.8%)	(15.6%)	(40.6%)
Visayan	2	1	3	6
	(6.25%)	(3.12%)	(9.37%)	(18.7%)
Total	10	9	13	32
	(31.3%)	(28.1%)	(40.6%)	(100%)

Table 6: Consolidated results of the senior high school learners' assessment on pronunciation of the English words with voiced /th/ sound

4 Discussion

From the data presented, it was found out that 19 out of 32 or 59.5% of the Grade 11 senior high school learner-participants of Enerdino C. Coronel – Baluno National High School were excellent in the pronunciation of the English words with unvoiced fricative consonant or the /θ/ sound. This means that majority of them pronounced the words with unvoiced /th/ sound almost perfectly, or at least committed at least one to three errors in the pronunciation of the English words in Set A minimal pairs reading aloud assessment. However, 11 out of 32 or 34.3% of the Grade 11 learner-participants were poor in pronouncing the English words with voiceless /th/ sounds. This means that they needed to exert more effort or practice more in the pronunciation of the words with /θ/ sound in their target language. Apparently, these 11 senior high school learner-participants committed seven to nine pronunciation errors, or possibly mispronounced all the English words with voiceless /th/ sounds in Set B minimal pair reading aloud test.

Conversely, only 10 out of 32 senior high school learner-participants or 31.3% were rated excellent in the pronunciation of English words with voiced fricative consonant or the /ð/ sound, while nine (9) of them or 28.1% were average, and 13 or 40.6% were poor in the pronunciation test. This signifies that majority of the Grade 11 learner-participants needed to improve their pronunciation of the words with voiced dental fricative sound in their target language. More than a third of the total sample size committed seven to nine pronunciation errors, or possibly mispronounced all the English words with voiced /th/ sound in Set B minimal pair reading aloud test.

In an ethnolinguistic perspective, majority of the Chavacano, Tausug, and Visayan senior high school learner-participants were excellent in the pronunciation of the words with unvoiced dental fricative or the /θ/ sound in their target language with 21.9% of the 13 speakers, 25.0% of the 13 speakers, and 12.6% of the 6 speakers respectively. This implies that most of the Grade 11 learner-participants did very well in pronouncing the English words with unvoiced /th/ sound in Set A minimal pair reading aloud test, projecting only minimal pronunciation errors. However, 15.6% of the Chavacano speakers, 15.6% of the Tausug speakers, and 3.1% of the Visayan speakers needed to intensely improve their pronunciation of the English words with unvoiced /th/ sound as they committed seven to nine pronunciation errors in the reading aloud assessment.

Moreover, although 18.8% of the 13 Chavacano speakers were excellent in the pronunciation of the words with voiced fricative consonant or the /ð/ sound, 15.6% of the Chavacano speakers were poor in the Set B minimal pair reading aloud test. This means that the Chavacano speakers tend to pronounce the English words with voiced fricative consonant more accurately than the other ethnolinguistic groups. Among the 13 Tausug speakers, 18.8% dominated the average level and 15.6% were poor. It implies that this ethnolinguistic group needed to practice more in their pronunciation of English words with voiced /th/ sound as they committed four to six and seven to nine pronunciation errors in the two levels respectively. There is also a possibility that the five (5) Tausug speakers never pronounced all the English words correctly in the Set B minimal pair reading aloud test. Finally, majority of the Visayan senior high school learner-participants or 9.37% of the 6 Visayan speakers were poor in the pronunciation of words with voiced dental fricative sound in their target language. Only 6.25% of the Visayan speakers were excellent, which means that this ethnolinguistic group tend to mispronounce the words with voiced /th/ sound more often than pronouncing them accurately.

In summary, most of the learner-participants in this study did very well in pronouncing the English words with /θ/ sound or the unvoiced fricative consonant but needed to improve their pronunciation of the words with /ð/ so und or the voiced fricative consonant. Among the three ethnolinguistic groups, the Tausug and Visayan speakers tend to pronounce the words with voiceless dental fricative more accurately than the Chavacano speakers. However, the Chavacano speakers tend to pronounce the words with voiced /th/ sound better than the Tausug and Visayan speakers.

Mispronunciation of the English words with fricative consonants made by the learner-participants are seen to be a manifestation of interlanguage error theory (Richards, 1974). This type of language error usually happens when the language learner makes a negative transfer of the linguistic elements of the native language (L1) into the target language (L2). In this case, the negative transfer made by the Chavacano, Tausug, and Visayan speakers was in the phonological aspect of the English language. As stated, the phonemic fricative consonant of /th/ either voiced or voiceless is not present in the Filipino language or in any local language in the Philippines, thus speakers disregard the pronunciation of words with voiced and voiceless dental fricative sound unless required in an English classroom instruction.

In addition, the geographic location was also a factor in the pronunciation accuracy of the learner-participants. Residents of Barangay Baluno, Zamboanga City only use Chavacano, Tausug, Visayan, or Filipino language in their usual conversations with their family, friends, or anyone they know. They would prefer to speak using the common local language instead of speaking

English, even in the actual English classes. Less exposure to the target language results to more chances of making a mistake in the L2.

Finally, may the findings of this study englighten the language instructors, the learners, and the English program supervisors to give proper instructions, practice, and interventions to improve the oral skill specifically the pronunciation of the learners. In the end, good pronunciation is a key element to an effective oral communication that gives a lasting impact to the listeners.

References

Adila, S. & Refnaldi, R. (2019). *Pronunciation errors made by senior high school students in speaking performance,* *8*(3). https://doi.org/10.24036/jelt.v8i3.105298

Almutalabi, M. (2018). Analysing the improper pronunciation of diphthongs by Iraqi EFL learners. *Jurnal Arbitrer*, *5*(1), 17-22. https://doi.10.25077/ar.5.1.17-22.2018

Creese, A. & Martin, P. (Eds.). (2003). Multilingual classroom ecologies: Inter-relationships, interactions and ideologies. Clevedon: Multilingual Matters.

Derwing, T. & Munro, M. (2010). Symposium – Accentuating the positive: directions in pronunciation research. *Language Teaching*, 43(*3*), 366-368. https://doi.org/10.10170S026144481000008X

Eckert, P. & Rickford, J. (2001). Style and sociolinguistic variation. Cambridge: Cambridge University Press.

Gilakjani, A. (2011). *Why is pronunciation so difficult to learn?* *4*(3). https://doi.10.5539/elt.v4n3p74

Haugen, E. (1972). *The ecology of language*. Stanford, CA: Stanford University Press.

Luna, J. & Sena, M. (2014). *Phoenix language package: Skill builders for English proficiency*. Phoenix Publishing House.

Keshavarz, M. H. (1999). *Contrastive analysis and error analysis.* Tehran: Rahnama Publication.

Lyons, D. (2021, March 10). How many people speak English, and where is it spoken? *Babbel Magazine.* https://www.babbel.com/en/magazine/how-many-people-speak-english-and-where-is-it-spoken#

Nagle, C., Levis, J., & Todey, E. (2019*)*. The changing face of L2 pronunciation research and teaching. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th pronunciation in Second Language Learning and Teaching Conference* (pp. 1-9). Iowa State University.

Pennington, M. C., & Rogerson-Revell, P. (2019). Relating pronunciation research and practice. *English pronunciation teaching and research*. Palgrave Macmillan London. https://doi.org/10.1057/978-1-137-47677-7

Richards, J. (1974). *Error analysis: Perspectives on second language acquisition.* London: Longman.

Szmigiera, M. (2022, April 1). The most spoken languages worldwide 2022. *Statista*. https://statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Tergujeff, E. (2012). English pronunciation teaching: Four case studies from Finland. *Journal of Language Teaching and Research*, *3*(4), 599-607.

Thomas, L. (2020). *Cross-sectional study | definitions, uses & examples.* https://scribbr.com/methodology/cross-sectional-study/

Yamaguchi, T. (2014). The pronunciation of TH in word-initial position in Malaysian English. *English Today, 30* (3), 13-21. https://doi.10.1017/S0266078414000224

SEND: A Simple and Efficient Noise Detection Algorithm for Vietnamese Real Estate Posts

Khanh Quoc Tran[1,3,4], **An Tran-Hoai Le**[1,3,4], **An Trong Nguyen**[1,3,4], **Tung Tran Nguyen Doan**[4], **Son Thanh Huynh**[2,3,4], **Bao Le Hoang**[2,3,4], **Hoang Nguyen Minh**[2,3,4], **Triet Minh Thai**[1,3,4], **Hoang Le Huy**[2,3,4], **Dang T. Huynh**[2,3,4], **Binh T. Nguyen**[2,3,4]*, **Nhi Ho**[5], **Trung T. Nguyen**[5]

[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *University of Science, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*
[4] *AISIA Research Lab, Vietnam*
[5] *Hung Thinh Corporation, Ho Chi Minh City, Vietnam*

Abstract

One of the emerging research fields in Natural Language Processing is Noise Detection (ND), the process of identifying posts containing noise information on textual data. While numerous datasets and approaches are developed for ND research in other languages, equivalent resources for the Vietnamese are limited. To the best of our knowledge, no dataset or method has been investigated or proposed to address the noise Detection tasks in the Vietnamese language. In reality, noise data is constantly present in datasets and sometimes hurts relevant model performance. To overcome this limitation, we propose ViND, a first human-annotated dataset that is available to the scientific community as a benchmark for the task of **Vi**etnamese **N**oise **D**etection. The ViND dataset contains 12,862 posts collected from five major Vietnamese real estate news websites. This paper provides an overview of the Vietnamese Noise Detection task, the process of creating the ViND dataset, and the techniques for carrying out the baseline experiments. On the ViND dataset, the PhoBERT$_{large}$ model outperforms robust baseline models such as LSTM, Bi-LSTM, BERT, RoBERTa, XLM-R, and DistilBERT and achieves a macro F1-score of 0.9024. In addition, our proposed method also successfully improves the related task's performance, mainly Vietnamese Named Entity Recognition (NER) for real estate posts, about 0.0239 in terms of macro F1-score.

1 Introduction

In Natural Language Processing (NLP) and Machine Learning, data and data processing play a significant role, especially when working with user-generated or social network content formed in non-standardized text. Furthermore, data in this aspect generally contains meaningless or useless, and this kind of information often impacts negatively on purpose but is easily ignored. Therefore, to increase the quality of data and the performance of the NLP models, removing all of the noisy information from the dataset (Subramaniam et al., 2009) is necessary.

The explosion of data available from social networking and e-commerce platforms has opened the need and opportunities for noise information processing in NLP (Al Sharou et al., 2021). However, for real estate, the posts could be disturbed by the wrong input from creators, missing essential descriptions, and the mistakes made by real estate agents. Besides, there is a wide range of helpful values with complete information. Nevertheless, their definitions are still ambiguous and complicated, making them difficult to distinguish from the actual noise (Subramaniam et al., 2009). Moreover, Vietnamese real estate post data also includes many challenges such as abbreviations, spelling errors, or some unclear situations that lead to misunderstandings for readers.

In this research, we present a study on building a Vietnamese real estate post dataset and a proposed method for Noise Detection in Vietnamese real estate posts data to improve the efficiency of other vital tasks on this data. Firstly, we collected data from Vietnam real estate from websites. Next, according to annotation guidelines, we annotated the data to noise or non-noise. Finally, we conduct experiments on noise classification methods to compare, analyze and propose a suitable technique. Three main contributions of this paper are summarized as follows:

*Corresponding author: Binh T. Nguyen (e-mail: ngtbinh@hcmus.edu.vn).

1. We present ViND, the first manually-annotated dataset created to serve as a benchmark for the task of Vietnamese Noise Detection. There are 12,862 posts annotated using a strict and efficient process to assure the dataset's quality.

2. Based on the best-performing model PhoBERT$_{large}$, we proposed a simple and efficient method for the Vietnamese Noise Detection task. Various experiments were implemented and evaluated on the ViND dataset using state-of-the-art baseline models. Moreover, we experimented with a combination of Noise Detection and the NER task for Vietnamese real estate posts to verify the effectiveness and contribution of the proposed method.

3. We have analyzed the error cases, limitations, and specific case studies that need to be addressed to improve the models' performance and develop further studies.

The rest of this paper can be organized as follows. First, in Section 2, we survey and describe an overview of the fundamentals of the Noise Detection task and relevant studies. Next, Section 3 shows the process of building our ViND dataset, including three stages: data collection, data annotation, and validation of annotation. Then, Section 4 contains our experiment and analysis on the ViND dataset, which includes the performance of the baseline models and common error cases. Finally, Section 5 provides our main conclusion and future works.

2 Fundamental of Noise Detection

2.1 Task Definition

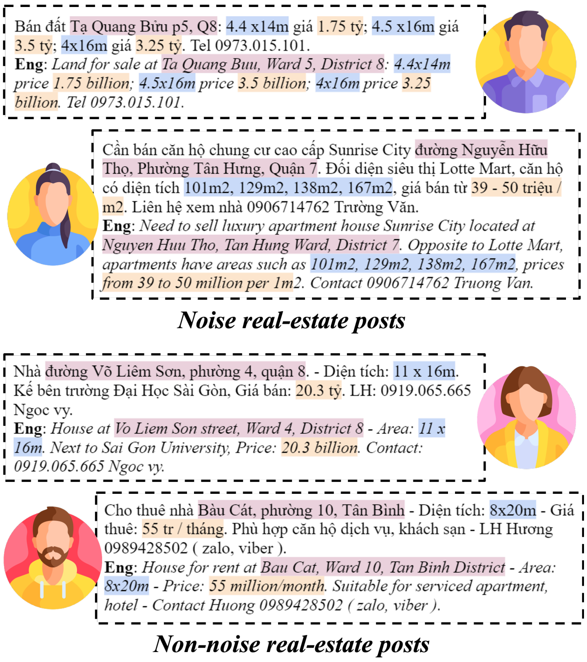
The starting step for the Noise Detection task is to determine a proper noise definition (Al Sharou et al., 2021). It is worth noting that the meaning of the noise can be different on diverse issues. In this section, we aim to recapitulate the Vietnamese Noise Detection task. The goal of this task is to classify the label **y** (noise or non-noise) corresponding to a provided real-estate advertisement post **X**.

Input: Given Vietnamese real estate posts on the real estate websites.

Output: One of the two labels described below.

1. **Noisy real estate posts (NOISE)** contains noisy data that are frequently intended to cause confusion and can harm the impression of information about a certain real estate. A post is identified as NOISE if it (1) mention many real estates in a single post; (2) does not provide critical information such as an address, price, or area; (3) refers to numerous pricing and places for one real estate.

2. **Non-noise real estate posts (Non-NOISE)** is a normal post. It can be an advertisement, brokerage, buying, or selling of real estate that contains transparent and necessary information without being confusing or restrictive.



Bán đất Tạ Quang Bửu p5, Q8: 4.4 x14m giá 1.75 tỷ; 4.5 x16m giá 3.5 tỷ; 4x16m giá 3.25 tỷ. Tel 0973.015.101.
Eng: *Land for sale at Ta Quang Buu, Ward 5, District 8: 4.4x14m price 1.75 billion; 4.5x16m price 3.5 billion; 4x16m price 3.25 billion. Tel 0973.015.101.*

Cần bán căn hộ chung cư cao cấp Sunrise City đường Nguyễn Hữu Thọ, Phường Tân Hưng, Quận 7. Đối diện siêu thị Lotte Mart, căn hộ có diện tích 101m2, 129m2, 138m2, 167m2, giá bán từ 39 - 50 triệu / m2. Liên hệ xem nhà 0906714762 Trường Văn.
Eng: *Need to sell luxury apartment house Sunrise City located at Nguyen Huu Tho, Tan Hung Ward, District 7. Opposite to Lotte Mart, apartments have areas such as 101m2, 129m2, 138m2, 167m2, prices from 39 to 50 million per 1m2. Contact 0906714762 Truong Van.*

Noise real-estate posts

Nhà đường Võ Liêm Sơn, phường 4, quận 8. - Diện tích: 11 x 16m. Kế bên trường Đại Học Sài Gòn, Giá bán: 20.3 tỷ. LH: 0919.065.665 Ngoc vy.
Eng: *House at Vo Liem Son street, Ward 4, District 8 - Area: 11 x 16m. Next to Sai Gon University, Price: 20.3 billion. Contact: 0919.065.665 Ngoc vy.*

Cho thuê nhà Bàu Cát, phường 10, Tân Bình - Diện tích: 8x20m - Giá thuê: 55 tr / tháng. Phù hợp căn hộ dịch vụ, khách sạn - LH Hương 0989428502 (zalo, viber).
Eng: *House for rent at Bau Cat, Ward 10, Tan Binh District - Area: 8x20m - Price: 55 million/month. Suitable for serviced apartment, hotel - Contact Huong 0989428502 (zalo, viber).*

Non-noise real-estate posts

Figure 1: Several instances of the Noise Detection task in Vietnamese.

2.2 Existing Methods for Noise Detection

Subramaniam et al. (Subramaniam et al., 2009) presented a picture of text noise types for documents. The authors surveyed different research topics, including Information Retrieval, Text Classification, Text Summarization, and Information Extraction tasks. The review showed general text noises for different sources of documents from different task aspects. Therefore, there would be many methods for each text noise type of the noise-detection task. Jindal et al. (Jindal et al., 2019) proposed a framework to enable a DNN to learn better sentence representations in the presence of label noise for text classification tasks.

It helped noise models absorb most of the label noise. Bagla et al. (Kumar et al., 2020) conducted experiments by using SOTA methods such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc., on IMDB movie reviews datasets (Maas et al., 2011) and Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) for the binary text classification task. The research also pointed out the importance of being mindful of any noise in text data when fine-tuning NLP models on noise text data.

2.3 Noise Detection in Vietnamese Real Estate

Text-noise is observable in most digital texts, including emails, SMS, blogs, and so on (Subramaniam et al., 2009). Furthermore, as real estate post data can be contributed by the community, being affected by noise is unavoidable. Noisy texts will degrade the performance of machine learning models, particularly transformer-based models (Kumar et al., 2020). Despite significant global development in this field (Jindal et al., 2019; Kumar et al., 2020), research in Vietnam is still limited. According to our survey, there is no research on this topic in Vietnam. As a result, we propose implementing Noise Detection into real estate posts. Our work contributes to the first dataset for Vietnamese Noise Detection in real estate posts. It also implements SOTA methods such as deep neural networks and transformer-based models to evaluate the findings.

3 Dataset Creation

Figure 2 depicts an overview of the process we performed to make the ViND dataset. Our dataset creation process goes through three phases: Dataset Collection, Data Annotation, and Validation of Annotation. These phases are described in detail as follows.

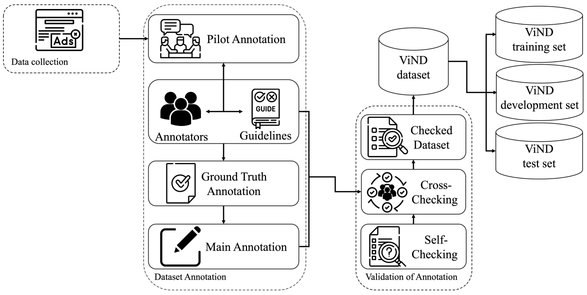


Figure 2: The procedure of creating ViND dataset.

3.1 Data Collection

Our data for this study comes from publicly available online sources, including major real estate websites in Vietnam, including https://nhadat247.com.vn, www.prozy.vn, https://homedy.com, https://muaban.net, and www.batdongsan.com.vn. For data collection, we use the robust Python tool - Beautiful Soup[1]. This library significantly assists us in extracting data from HTML or XML files by providing Pythonic techniques for iterating, searching, or editing directly on the parse tree. After the crawling is completed, an appropriate database is constructed to store the data.

Each collected sample of real estate posts typically includes a post description and any additional information such as an address, area, and price. In some situations, such as when the post description contains undesirable artifacts, erroneous characters, or HTML markers, it is still possible to collect this information. Nonetheless, raw data could occasionally lack critical information from post descriptions. Therefore, combining all relevant features collected from the post description with those already available in raw data could be critical in building an effective data collection strategy from multiple sources. In the end, we collected 12,862 samples to create the ViND dataset.

3.2 Data Annotation Process

Metric For Inter-Annotator Agreement Cohen's Kappa is commonly used to evaluate inter-annotator agreement (IAA) in several tasks and is widely considered as the benchmark (McHugh, 2012). As a result, we employ Cohen's Kappa (Bhowmick et al., 2008) to compute inter-annotator agreements of annotators and quality assurance of human annotation. The Cohen's Kappa coefficient can be formulated as follows:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{1}$$

where k represents an inter-annotator agreement, $Pr(a)$ represents a relative observed agreement among raters, and $Pr(e)$ represents the hypothetical probability of chance agreement.

3.2.1 Stage 1: Pilot Annotation

We recruited six undergraduate students for our annotation tasks. Most of them had experience

[1] https://pypi.org/project/beautifulsoup4/

annotating several datasets in Vietnamese Natural Language Processing. The primary goal of this Pilot Annotation Stage was to acquaint our annotators with the task. After that, we created an initial set of annotation guidelines with examples and sent them to annotators. Then, before annotating the same 200 random samples from the collected data, all annotators were asked to proofread carefully and rigorously adhere to the annotation guidelines. They repeated these steps five times to compute IAA using Cohen's Kappa, which was obtained by averaging the results of pairwise comparisons among all annotators. Finally, annotators annotated the data independently until the inter-annotator agreement of the annotations achieved more than 0.80 (near perfect agreement) (McHugh, 2012), and they completed the annotation guidelines. Figure 3 presents the IAA of our staff on the tasks of Vietnamese Noise Detection during the Pilot Annotation stage.

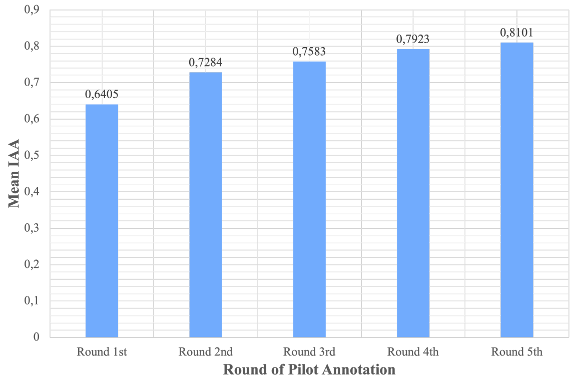


Figure 3: Inter-annotator agreements from five separate annotation training rounds.

Annotation Guidelines A detailed annotation guideline has been explicitly composed for the annotators to identify and label noises efficiently. The proposed guideline includes the following phases: (1) If the annotating post contains sufficient information in the description, it will be labeled as not noise; (2) Any post whose description mentions more than one real estate or contains many values for the same property, such as area or price, can be labeled as noise; (3) There are also posts with either selling or renting types or a combination of these two types that do not provide any information about the address and area or address and price of a specific real estate. Again, it implies that those posts can be labeled "noise" when no helpful information is found.

Sometimes the guidelines do not cover all the diversity of real estate posts and can create challenges that the annotation team may face during the annotation process. Some rare cases where the post contains ambiguous information that is too difficult to identify noise can be set aside for later discussion. After each discussion, the guideline will be modified in detail and complemented to generalize better to new cases in the future.

3.2.2 Stage 2: Ground Truth Annotation

We randomly selected a Ground Truth set of 1,200 posts from the collected data for this stage. Two guideline developers independently annotated the Ground Truth set using the well-developed guidelines from the previous step and reached an IAA of 0.87. These annotators have a deep awareness of the data and tutorials that ensure the reliability and efficacy of the annotation process. In addition, they discussed annotation concerns and solutions for further improvement.

3.2.3 Stage 3: Main Annotation

We divided the collected posts into six equal and non-duplicating subsets. Each well-trained annotation from Phase 1 will be assigned a subset to annotate. In addition, we add 200 samples from the ground truth set to each subset at random and separately. Annotators were asked to modify tasks until the IAA reached 0.80 or higher. Then, the IAA (Cohen's Kappa) was evaluated by comparing each annotator to the corresponding ground truth. This procedure was completed with a mean Cohen's Kappa of 0.83.

3.3 Validation of Annotations

We carefully validated the annotated data before publishing it for research purposes. We required our annotators to self-check the posts they had annotated and prepare short notes to report on their own mistakes after annotating every 500 samples to improve their annotation. This effort decreases the possibility of our annotators making the same mistake too often. To reduce the error rate, we have an additional step of cross-checking once we complete annotating every 3,000 samples. Our staff then investigates and validates any mistakes discovered by others.

3.4 Dataset Statistics

The ViND dataset contains 12,862 posts divided into three subsets: training, development, and test, with a ratio of 6:2:2. Basic statistics of the three ViND subsets are shown in Table 1[2]. We can see that the length of posts ranges from 8 to 1,038 words, with an average of 110 words utilized in each post. Besides, having a large quantity of training vocabulary allows us to train and fine-tune the models more effectively. Interestingly, posts labeled with noise are, on average, 12.2±1.5 words less in length than non-noise posts. It is because non-noise data contains posts that do not provide essential information, making it shorter.

Table 1: Basic statistics of proposed ViND dataset.

		Training set	Development set	Test set
Full Data	Number of posts	7,717	2,571	2,574
	Average posts length	109.5	111.2	108.3
	Total Vocabulary size	845,632	283,404	278,808
	Maximum posts length	1038	622	622
	Minimum posts length	8	11	8
Noise	Number of posts	1,524	417	420
	Average posts length	98.1	101.2	98.4
	Total Vocabulary size	123,117	42,229	41,337
	Maximum posts length	585	579	505
	Minimum posts length	8	13	8
Non-noise	Number of posts	6,463	2,154	2,154
	Average posts length	111.8	111.9	110.2
	Total Vocabulary size	722,515	241,175	237,471
	Maximum posts length	1038	622	622
	Minimum posts length	8	11	10

Figure 4 depicts the distribution statistics of the number of posts with and without noise labels in each ViND subset. We can observe from the statistical results that the dataset is unbalanced since posts with noise labels account for a considerable proportion of the total. It can be explained by the fact that, in real life, the real estate websites we choose to collect have a team of administrators who filter and eliminate noise posts. However, for objective reasons, a small proportion of such posts still exist and should be deleted from the system.

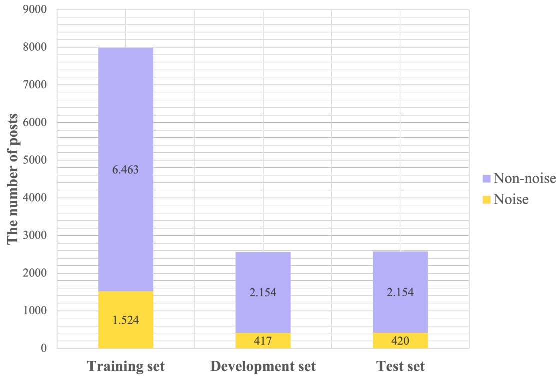


Figure 4: The labels distribution on each ViND subset.

4 Experiments

Figure 5 presents an overview of the experimental procedure for our task in this paper.

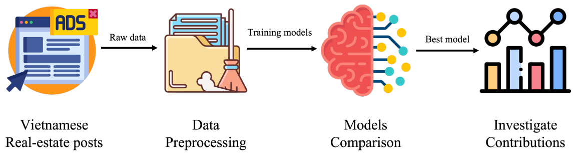


Figure 5: Overview of the experimental procedure for Vietnamese Noise Detection.

4.1 Data Preprocessing

The dataset is processed using the following techniques before it is used in experiments: (1) Normalizing text to Unicode standard; (2) Cleaning input formats (e.g., HTML or Javascript from crawlers); (3) Removing invalid characters (e.g., emojis, non-Vietnamese characters). **Example:** *Nhà đẹp lung linh*$_{*\textit{beautiful house}*}$ → Nhà đẹp lung linh$_{\textit{beautiful house}}$; (4) Fixing non-standard diacritical marks and non-standard punctuations. **Example:** Thanh Hoá → Thanh Hóa, uỷ ban$_{\textit{committee}}$ → ủy ban$_{\textit{committee}}$; (5) Using the VnCoreNLP tool (Vu et al., 2018) to do word segmentation.

4.2 Baseline models for Noise Detection performance comparison

We experiment with various approaches to classify noise data, including transformer-based pre-trained language models and deep neural network models. In this study, state-of-the-art models, including LSTM, Bi-LSTM, BERT, RoBERTa, XLM-R, DistilBERT, and PhoBERT, are implemented and fine-tuned to find the best model for the task of Vietnamese Noise Detection.

Long Short Term Memory (LSTM) is a particular type of RNN, which was introduced by Hochreiter et al. (Hochreiter and Schmidhuber, 1997) in 1997. This method has an additional memory cell compared to traditional RNN to capture long-term dependencies information. Moreover, LSTM also presents new gates like input and forget gates to control gradient flow, helping us avoid vanishing gradients when training.

Bidirectional-LSTM (Bi-LSTM) is an extended version of LSTM (Hochreiter and Schmidhuber, 1997) that was proposed by Grave et al. (Schuster and Paliwal, 1997). Unlike standard LSTM, this approach utilizes the information flow from both directions, thus, enhancing the understandability

[2] Note that vocabulary size and comment length are computed at the word level.

of the model. BiLSTM can be used in various NLP tasks like machine translation, named entity recognition, and our text classification task.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based approach for pre-trained in various NLP task which was developed by Delvin et al. in 2018. This technique was pre-trained on two tasks: masked language modeling and next sentence prediction. This work uses the training set for fine-tuning the pre-trained BERT model before classifying posts.

Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019) is trained with the help of dynamic masking. This technique forces the system to predict the hidden sections of text truly while unannotated language examples. RoBERTa, implemented in PyTorch, is an extension of the BERT approach (Devlin et al., 2019) with some adjustments in terms of key hyper-parameters like mini-batches size and learning rates. This method also discards the next-sentence pre-trained objective from the original BERT.

XLM-RoBERTa (XLM-R) (Conneau et al., 2020) is a multilingual language model introduced Conneau el at. in 2019. It is a variant of RoBERTa (Liu et al., 2019) which was pre-trained on 2.5T of data across 100 languages containing 137GB of Vietnamese texts. On several cross-lingual benchmarks, this technique outperforms mBERT (Conneau et al., 2020).

DistilBERT (Sanh et al., 2019) is a smaller version of BERT approach (Devlin et al., 2019) which was introduced by Sanh et al. in 2019. Although it just contains 40% fewer parameters than BERT, this method enables it to preserve over 95% of BERT's performance and execute 60% faster. Taking advantage of this efficient method, we utilize DistillBERT in our experiment with the belief of achieving a sustainable result.

PhoBERT (Nguyen and Tuan Nguyen, 2020) is a monolingual pre-trained language model that was trained on a 20GB Vietnamese dataset using the same architecture and approach as RoBERTa (Liu et al., 2019). PhoBERT enables to outperform many state-of-the-art approaches in Vietnamese-specific NLP tasks, including text classification (Nguyen and Tuan Nguyen, 2020; Tran et al., 2022, 2021).

4.3 Experimental Settings

We use the training set to fit experimental parameters and the development set to fine-tune classifier hyper-parameters. We utilize the test set to evaluate our baselines and implement an Adam optimizer (Kingma and Ba, 2015) with a Dropout of 0.2 and a fixed learning rate of 1e-5, and num_train_epochs equal to 4 to fine-tune the hyper-parameters of baseline models. Deep neural network models, including LSTM and Bi-LSTM, are implemented with a 300 embedding size and a Dense two output layer with a Sigmoid activation function. We also pass our input through several well-known word embedding (Tuan Nguyen et al., 2020; Vu Xuan et al., 2019) before feeding it to LSTM and Bi-LSTM model.

In this work, we use simpletransformers[3] to implement all pre-trained language models in transformer-based models. These pre-trained models have a max sequence length equal to 100 and a learning rate decay of 0.01. In addition, we also apply both variants of BERT, RoBERTa, XLM-R, and PhoBERT, including base and large versions.

4.4 Noise Detection Performance Evaluation Metrics

This section presents the evaluation metrics employed in this study. The commonly used metrics for text classification tasks in general (Sokolova and Lapalme, 2009), and detecting noise posts in particular, are Precision, Recall, and F1-score. However, because the proposed datasets have notably imbalanced classes, the average macro F1-score, the harmonic mean of Precision and Recall, is the optimal metric for this task. As a result, we used the average macro F1-score as the critical measure, with the Precision and Recall providing additional information.

4.5 Experimental Results

Table 2 shows our results from the experiments. Compared to deep learning models, the combination between Bi-LSTM and FastText has the highest F1 score of 0.6594 for the ViND test Full Data. Furthermore, the model mentioned above has an F1 score of 0.4119, the highest in the Noises Data category. Besides, combining Bi-LSTM with PhoW2V$_{word}$ achieves the greatest F1-score of 0.9077 in the Non-noises Data category.

[3] https://simpletransformers.ai/

Table 2: Noise Detection results on the ViND test set using various methods. The best outcomes in each category are highlighted.

Model	Full Data			Noises Data			Non-noises Data		
	Pre.	**Rec.**	**F1-score**	**Pre.**	**Rec.**	**F1-score**	**Pre.**	**Rec.**	**F1-score**
LSTM + fastText	0,8489	0,5972	0,6248	0,8333	0,2024	0,3257	0,8645	0,9921	0,9239
LSTM + Word2vec	0,8133	0,6116	0,6429	0,7576	0,2381	0,3623	0,8690	0,9851	0,9234
LSTM + $PhoW2V_{word}$	0,8449	0,6015	0,6307	0,8241	0,2119	0,3371	0,8658	0,9912	0,9242
LSTM + $PhoW2V_{syllable}$	0,8634	0,5946	0,6213	0,8632	0,1952	0,3184	0,8637	**0,9940**	0,9242
Bi-LSTM + fastText	0,6952	0,6404	0,6594	0,5106	0,3452	0,4119	0,8799	0,9355	0,9068
Bi-LSTM + Word2vec	0,6839	0,6383	0,6549	0,4883	0,3476	0,4061	0,8796	0,9290	0,9036
Bi-LSTM + $PhoW2V_{word}$	0,6976	0,6372	0,6573	0,5165	0,3357	0,4069	0,8787	0,9387	0,9077
Bi-LSTM + $PhoW2V_{syllable}$	0,6926	0,6406	0,6590	0,5052	0,3476	0,4118	0,8801	0,9336	0,9061
BERT$_{base}$	0,9172	0,8326	0,8674	0,8938	0,6810	0,7730	0,9406	0,9842	0,9619
BERT$_{large}$	0,8973	0,8701	0,8829	0,8390	0,7690	0,8025	0,9557	0,9712	0,9634
RoBERTa$_{base}$	0,9219	0,8209	0,8608	0,9076	0,6548	0,7607	0,9362	0,9870	0,9609
RoBERTa$_{large}$	0,9016	0,8820	0,8914	0,8430	0,7929	0,8172	0,9601	0,9712	0,9656
XLM-R$_{base}$	0,9307	0,8313	0,8711	0,9218	0,6738	0,7785	0,9396	0,9889	0,9636
XLM-R$_{large}$	0,9324	0,8270	0,8686	0,9269	0,6643	0,7739	0,9380	0,9898	0,9632
DistilBERT	0,9153	**0,8894**	0,9016	0,8686	**0,8024**	0,8342	**0,9620**	0,9763	0,9691
PhoBERT$_{base}$	**0,9416**	0,8661	0,8983	**0,9313**	0,7429	0,8265	0,9518	0,9893	0,9702
PhoBERT$_{large}$	0,9312	0,8790	**0,9024**	0,9053	0,7738	**0,8344**	0,9571	0,9842	**0,9705**

With transformer-based models, we get better outcomes in every area when utilizing DistilBERT and PhoBERT. For example, whereas DistilBERT has the two highest Recall scores for Full Data and Noises Data categories (0.8894 and 0.8024, respectively) and the highest value of 0.9620 for Precision of Non-noise Data category, PhoBERT$_{base}$ has the two highest Precision of Full Data and Noise Data with 0.9416 and 0.9313 orderly.

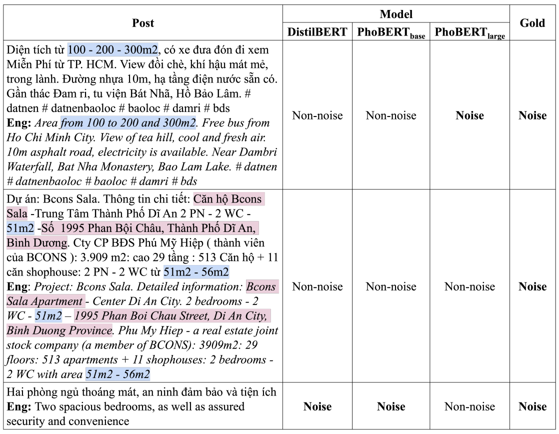
Experimental results indicate that the PhoBERT$_{large}$ model outperforms transformer-based models on Full Data, Noise Data, and Non-noise Data, respectively, by an F1 score of 0.9024, 0.8344, and 0.9705. The PhoBERT model, particularly the PhoBERT$_{large}$ model, has the benefit of being trained on a substantial Vietnamese data domain gathered from various subjects and news websites (Nguyen and Tuan Nguyen, 2020). Due to the diversity of language in the pre-trained and training data, domain knowledge and terminology are better represented, improving model performance.

Furthermore, the traditional technique using LSTM and PhoW2Vec provides an outstanding result in 0.9940 of Recall. Besides, one can see that almost the metric values of transformer-based models are higher than the older ones. On the other hand, the monolingual pre-trained language model for Vietnamese (PhoBERT) beats the multilingual models on the task of Vietnamese Noise Detection. It turns out that existing solutions can solve Noise Detection tasks and generate positive results.

4.6 Error Analysis and Discussion

Based on the macro F1 scores, we select the best baseline models, DistilBERT, PhoBERT$_{base}$, and PhoBERT$_{large}$, to perform error analysis. Then, as shown in Figure 6, we report the statistics of the ratio of various types of error cases[4] of 200 random samples in the ViND development set.

Table 3: Case studies in ViND development set. We evaluate DistilBERT, PhoBERT$_{base}$, and PhoBERT$_{large}$ on 3 sampled posts, with their gold labels and model predictions.



Post	Model			Gold
	DistilBERT	**PhoBERT$_{base}$**	**PhoBERT$_{large}$**	
Diện tích từ 100 - 200 - 300m2, có xe đưa đón đi xem Miễn Phí từ TP. HCM. View đồi chè, khí hậu mát mẻ, trong lành. Đường nhựa 10m, hạ tầng điện nước sẵn có. Gần thác Đam ri, tu viện Bát Nhã, Hồ Bảo Lâm. # datnen # datnenbaoloc # baoloc # damri # bds **Eng:** *Area from 100 to 200 and 300m2. Free bus from Ho Chi Minh City. View of tea hill, cool and fresh air. 10m asphalt road, electricity is available. Near Dambri Waterfall, Bat Nha Monastery, Bao Lam Lake. # datnen # datnenbaoloc # baoloc # damri # bds*	Non-noise	Non-noise	**Noise**	**Noise**
Dự án: Bcons Sala. Thông tin chi tiết: Căn hộ Bcons Sala -Trung Tâm Thành Phố Dĩ An 2 PN - 2 WC - 51m2 -Số 1995 Phan Bội Châu, Thành Phố Dĩ An, Bình Dương. Cty CP BĐS Phú Mỹ Hiệp (thành viên của BCONS): 3.909 m2: cao 29 tầng : 513 Căn hộ + 11 căn shophouse: 2 PN - 2 WC từ 51m2 - 56m2 **Eng**: *Project: Bcons Sala. Detailed information: Bcons Sala Apartment - Center Di An City. 2 bedrooms - 2 WC - 51m2 – 1995 Phan Boi Chau Street, Di An City, Binh Duong Province. Phu My Hiep - a real estate joint stock company (a member of BCONS): 3909m2: 29 floors: 513 apartments + 11 shophouses: 2 bedrooms - 2 WC with area 51m2 - 56m2*	Non-noise	Non-noise	Non-noise	**Noise**
Hai phòng ngủ thoáng mát, an ninh đảm bảo và tiện ích **Eng:** Two spacious bedrooms, as well as assured security and convenience	**Noise**	**Noise**	Non-noise	**Noise**

We notice that *Information overlap*[5], *Needing syntactic knowledge*[6], and *Short sequence*[7] are

[4] Definition of errors are explained in the Appendix A.

[5] The presence of several overlapping items of information on a given property in the case.

[6] The case contains complicated syntactic structure, and the model is unable to recognize the exact meaning.

[7] The input case is very short (< 20 words).

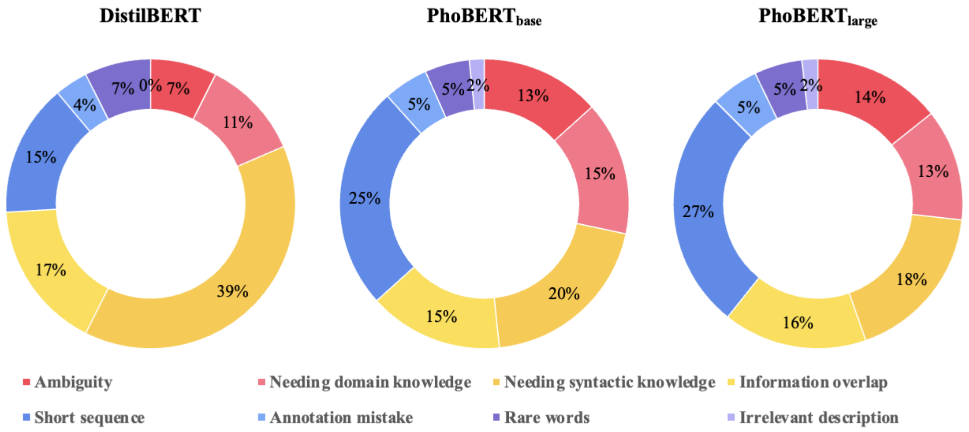


Figure 6: We conduct error analysis on ViND datasets with three best-performing models. We divide error cases into eight categories: Ambiguity, Needing domain knowledge, Needing syntactic knowledge, Information overlap, Short sequence, Annotation mistake, Rare words, and Irrelevant description.

the most common types of failure in DistilBERT, PhoBERT$_{base}$, and PhoBERT$_{large}$. DistilBERT model, in particular, has a high proportion of type Needing domain knowledge errors because it was pre-trained on a limited Vietnamese data domain.

We further show some cases study on ViND development set in Table 3. In the first case, we notice that both DistilBERT and PhoBERT$_{base}$ fail to predict the post's label, while PhoBERT$_{large}$ can obtain the correct prediction. Since this post includes real estate terms and PhoBERT$_{large}$ pre-trained language models that have the advantage of being trained on a larger Vietnamese data domain, one could use domain knowledge to comprehend those terms. In the second case, all three models fail to detect the noise since the post is highly complicated and contains information overlap. In the last point, PhoBERT$_{large}$ incorrectly predicts a short post because the model requires a significant amount of information to identify.

As depicted in Figure 6 and Table 3, overlapping information, lack of information due to limited input data, and complexity in posting structure are challenging issues that need to be addressed in the future. In conclusion, we conclude that the task of Vietnamese Noise Detection is challenging due to the unique peculiarities of the Vietnamese language, and more state-of-the-art methods should be investigated and proposed.

4.7 Contribution Verification For The Proposed Noise Detection Model

We have investigated the proposed method's effectiveness in reducing noise posts in the NER task for Vietnamese real estate posts. In experiments, we utilize our best model, PhoBERT$_{large}$, as a classifier, and the system can eliminate the advertisement posts predicted by the model to be noise. In contrast, the remaining posts can be used to train and evaluate NER models.

To ensure reliability and transparency in the comparison, we conduct the experiment using the same data partitioning, experimental settings, and metrics as Son et al. (Huynh et al., 2021). Furthermore, we decide to compare our approach with Son et al. (2021) as it is the most recent research in the Vietnamese NER task and is in the same field as the real estate news we are addressing.

The experimental results are presented in Tale 4. The results show that our proposed Noise Detection method significantly improves the NER performance (increase up to 0.0239 F1-score). As a result, using PhoBERT$_{large}$ to reduce noise data is efficient and generates state-of-the-art results on the task of NER for Vietnamese real estate.

5 Conclusion and Future Works

This paper presents ViND, a new Vietnamese real estate posts dataset for Noise Detection, including 12,862 samples. In addition, we implement LSTM and BERT models to find a better model that achieves better performance in the Noise Detection task. Finally, from the obtained results, we analyze and record typical error cases that need to be handled to help improve model performance.

As discussed in Section 4.6, we will investigate and test approaches for dealing with imbalanced, overlapping data to improve the solution's performance in the future. Moreover, our research lays

Table 4: The results compare NER models performance without and with using Noise Detection.

Model	Previous study (Huynh et al., 2021)			Our Noise Detection + (Huynh et al., 2021)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
MaxoutWindowEncoder W64	0.8623	0.8933	0.8775	0.8739 (↑0.0116)	0.9224 (↑0.0291)	0.8975 (↑0.0200)
LSTM W64	0.8486	0.8628	0.8556	0.8581 (↑0.0095)	0.8591 (↓0.0037)	0.8586 (↑0.0030)
MishtWindowEncoder W64	0.8677	0.8669	0.8673	0.8753 (↑0.0076)	0.8874 (↑0.0205)	0.8813 (↑0.0140)
BiLSTM W64	0.8573	0.8331	0.8450	0.8547 (↓0.0026)	0.8525 (↑0.0194)	0.8536 (↑0.0086)
MaxoutWindowEncoder W300	0.8739	0.8871	0.8805	0.8783 (↑0.0044)	0.8968 (↑0.0097)	0.8875 (↑0.0070)
LSTM W300	0.8649	0.8869	0.8758	0.8675 (↑0.0026)	0.8946 (↑0.0077)	0.8808 (↑0.0050)
MishWindowEncoder W300	0.8914	0.9237	0.9072	0.9174 (↑0.0260)	0.9452 (↑0.0215)	0.9311 (↑0.0239)
BILSTM W300	0.8524	0.8549	0.8535	0.8725 (↑0.0201)	0.8782 (↑0.0233)	0.8753 (↑0.0218)

the groundwork for various emerging research in Natural Language Processing, such as: (1) Named entity recognition; (2) Text classification; (3) Natural Language Inference.

Limitations and Ethics

We recognize the risk of releasing a dataset for detecting noise texts. For example, because of the subjectivity of manual annotation, our dataset may contain mislabeled data. In addition, due to limits in data coverage and training approaches, our benchmarks cannot detect all types of noise. However, we believe that our proposed benchmark provides more advantages than risks.

All comments in ViND are collected from real estate news websites. This study has ensured user anonymity by eliminating all relevant information when constructing the dataset and rigorously adhering to data source protocol. As a result, the items in our dataset **DO NOT** reflect our opinions or thoughts. ViND is made available to the public for research purposes only.

Acknowledgments

We thank the University of Science, Vietnam National University Ho Chi Minh City, Hung Thinh Corp., and AISIA Research Lab for supporting us throughout this paper. This research is funded by Hung Thinh Corp. under grant number HTHT2021-18-01.

References

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Plaban Kumar Bhowmick, Anupam Basu, and Pabitra Mitra. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Son Huynh, Khiem Le, Nhi Dang, Bao Le, Dang Huynh, Binh T. Nguyen, Trung T. Nguyen, and Nhi Y. T. Ho. 2021. Named entity recognition for vietnamese real estate advertisements. In *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 23–28.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy text data: Achilles' heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

L. Venkata Subramaniam, Shourya Roy, Tanveer A. Faruquie, and Sumit Negi. 2009. A survey of types of text noise and techniques to handle noisy text. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09, page 115–122, New York, NY, USA. Association for Computing Machinery.

Khanh Q. Tran, An T. Nguyen, Phu Gia Hoang, Canh Duc Luu, Trong-Hop Do, and Kiet Van Nguyen. 2022. Vietnamese hate and offensive detection using phobert-cnn and social media streaming data.

Khanh Quoc Tran, Phap Ngoc Trinh, Khoa Nguyen-Anh Tran, An Tran-Hoai Le, Luan Van Ha, and Kiet Van Nguyen. 2021. An empirical investigation of online news classification on an open-domain, large-scale and high-quality dataset in vietnamese. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 367–379. IOS Press.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60.

Son Vu Xuan, Thanh Vu, Son Tran, and Lili Jiang. 2019. ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1285–1294.

A Definition of Error Cases for Error Analysis

We introduce the error definition as follows and illustrate some error cases for Vietnamese Noise Detection tasks in Figure 6:

- **Ambiguity:** the case has the same context but a distinct meaning, which causes the prediction to be incorrect.
- **Needing domain knowledge:** there is real-estate terminology in the case that requires domain knowledge to comprehend.
- **Needing syntactic knowledge:** the case contains complicated syntactic structure, and the model fails to recognize the exact meaning.
- **Information overlap:** the presence of several overlapping items of information on a given property in the case.
- **Short sequence:** the input case is very short (< 20 words).
- **Annotation mistake:** the annotated label is incorrect.
- **Rare words:** the case contains low-frequency terms.
- **Irrelevant description:** the instance has a large amount of irrelevant information, which causes the prediction to be incorrect.

Gunita: Visualizing the evolution of the Philippines' languages with a historical Philippine text corpora

Amiel Bornales
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
amiel_bornales@dlsu.edu.ph

Jonn Yuvallos
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
jonn_yuvallos@dlsu.edu.ph

Courtney Ngo
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
courtney.ngo@dlsu.edu.ph

Abstract

While there are many culturomic studies in other countries, only a few studies focus on the culturomics unique to the Philippines. This study developed a Philippine news sources scraper and used a pre-existing Tagalog corpora containing books and poems across 100 years to build a continuously growing corpus. This study introduces Gunita, a web application that allows users to visualize how an n-gram is used over time and know which article, book, or poem the n-gram is used in to shed light on how Filipinos communicate through written text.

1 Introduction

Culture is an ever-evolving aspect of society. It provides each person their own cultural identities and heritages, and is an essential area of human society that is ripe with research. This field of research is known for its cultural studies, and it aims to explore the connections between gender, race, class, etc. and their effects on culture (Barker, 2003). While cultural studies are able to draw powerful conclusions on human culture, these conclusions are based off of a collection of carefully chosen works that represent only a minority of the media available at the time (Michel et al., 2011). Analysis of cultural trends has always been hampered by the lack of suitable data, and so, to help further research along, a corpus containing 5,195,769 digitized books was created and analyzed (Michel et al., 2011). This led to the creation of the field of culturomics, which is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts.

2 Related Works

The analysis of language and words is an important aspect of culturomics, and large text corpora has been a necessity for language analysis for many of these studies. However, these researches differ in how they handle their data. For example, Michel et al. (2011) used a large corpus to investigate cultural trends between 1800 and 2000, such as insights into lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Given that their scope is extremely broad, they focused more on the general state of things, and only go into specifics for important figures or events. For example, it was shown that Tiananmen (written in Chinese characters) was barely mentioned while the English term "Tiananmen" was mentioned increasingly in this period. This shows that the Chinese term of "Tiananmen" was not heavily used in China after the massacre due to censorship, while the English term of "Tiananmen" was mentioned frequently in

Western media during the aftermath. This research will adopt the same focus and analyze the data based on news and historical events.

According to Leetaru (2011), it is possible to analyze cultural trends even further with culturomics because it seeks to provide insights and encourage explorations on broad cultural trends. This is done by analyzing vast digital corpora computationally because the extremely large amount of textual data required for a proper corpus is impossible to be manually read by a human (Michel et al., 2011). Once the data has been processed, the data is reflective of the time the data came from, which gives us a snapshot of how the information environment was back then (Leetaru, 2011).

Ilao et al. (2011) performed culturomic analysis on online tabloid articles and was able to provide insights on various aspects of Filipino culture, like the Filipino Christmas tradition, the Filipino outlook on extreme calamities, and linguistic trends. With that in mind, this study also aims to study the cultural evolution of the Philippines, with a strong emphasis on the evolution of a few of the languages in the Philippines. These languages have changed significantly over the years, in terms of spelling, vocabulary, orthography, and grammar. In the case of Filipino, this is due to a variety of reasons, such as transitioning to a modern spelling system from Spanish-influenced spelling system, the release of the Grammar of the National Language, and the language wars that occurred when the Philippines was deciding on its national language (Ilao and Guevara, 2012). However, it is difficult to accurately trace the history of the other languages in the Philippines such as Cebuano and Ilocano due to the lack of available historical resources.

3 Methodology

The structure of the data input flow for this study is illustrated in Figure 1. It is divided into three stages: data collection, data processing, and data storage. After the data has been stored, the Gunita system can be used to gain insights on specific n-grams using the three features: n-gram usage visualization (line and frequency chart), wildcard search (simplified regular expressions), and co-occurrence search.

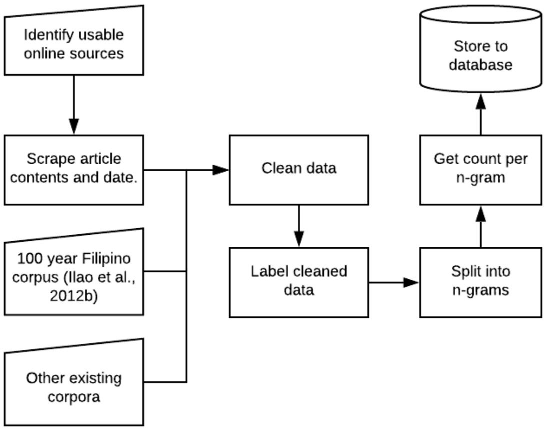

Identify usable online sources
Scrape article contents and date.
100 year Filipino corpus (Ilao et al., 2012b)
Other existing corpora
Clean data
Label cleaned data
Split into n-grams
Get count per n-gram
Store to database

Figure 1. The data input flow this project comes from a Philippine news source scraper and from the text corpus of Ilao and Guevara (2012). The raw HTML data from the scraper is stripped off HTML tags and other characters, and the whole article is labelled a language using FastText. The result is split into n-grams, underwent n-gram counting, and stored into the database.

3.1 Data Collection

The data will come from two sources. The first source for the data is scraped from online news articles and websites, and the second source came from the historical Tagalog corpus by Ilao and Guevara (2012).

Before any scraping is done on a target source website, it must have a directory containing all articles, and the article must have a publishing date. If a website passes this criteria, Python's request library is used to send requests to the web server’s article directory and receive the HTML response. Then, the article links from the raw HTML data is retrieved with the help of a Python library called BeautifulSoup, which allows the querying of the article links using an HTML parser. Lastly, all the article links are saved in CSV file, with each target source having their own CSV file to avoid any confusion and accidental mixing of data. Once all the links have been collected, the crawler visits each article link and scrapes all its article text content. The text contents are also added as a separate column in its corresponding CSV file.

3.2 Language Labeling

To label each article by its language, the language classifier from FastText was used. However, since the model classifies 176 different languages, there are cases where the article that was written in one of the four given languages

(Tagalog, Cebuano, Ilocano, English) is labelled as some other language. To fix this, the researchers extracts the language probabilities returned by the model. A sample result would be ['english' : .90, 'tagalog' : .40, 'cebuano': .20], which means that the probability of the text being english is 90%, tagalog 40%, and cebuano 20%. The article will then be classified as the language that has the highest probability given that it is either Tagalog, Cebuano, Ilocano, or English.

3.3 Data Processing

From the raw gathered article data, BeautifulSoup's built-in HTML parser was used to remove the HTML tags, and Python's built-in regular expression (RegEx) library was used to remove unneccessary special symbols. The RegEx used performs two operations: the first is to remove any periods or commas that are not used in numbers, and second is to remove any quotation marks so there would be no issues when exporting the cleaned data into a CSV file. To determine whether the period or comma is used in a number, the RegEx library checks if its adjacent characters are numbers, for example “1,234” or “1.23m” will retain its period and commas, however the string “cat.dog” will have the period removed since it does not meet the requirement of having 2 adjacent numbers. The usage of capture groups in RegEx will be used in retaining the characters adjacent to the removed period or comma.

After cleaning the data, the words will be separated into n-grams, where n will range from 1 to 3. To create and count the n-grams, the CountVectorizer from the Python library sklearn was used.

3.4 Data Storage

Once the data has been cleaned it is stored into the database. The entity relationship diagram of the database is shown in Figure 2.

Database Tables: The ngram_occurrences table contains an auto incremental ngram_occ_id as the primary key, ngram_id as the foreign key referencing the ngram in the ngram master list, n_number as the size of the n-gram (1, 2, or 3), ngram as the n-gram itself, date as the date the n-gram was published, lng_src is the language source of the n-gram (1 for Cebuano, 2 for Tagalog, and 3 for Ilocano), date as the date the data was taken from, n_number as the total amount of occurrences the ngram has on that date, source_id and source_link are foreign keys that reference the source of the ngram.

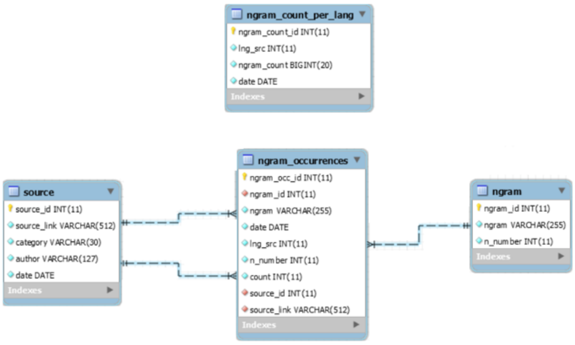


Figure 2. The database has a table for the source URIs, n-grams, n-gram counts for each language, and the n-gram occurrences which includes the URI the word was mentioned, the date the article was published, and how many times the word was mentioned in that article.

The database is also indexed on multiple columns, the index is ngram_occurrences_idx_ngram_src_number_date. This multi-column index contains the ngram, lng_src, n_number, and date as its indexes. This multi-column index was added to the database because a query can only make use of one index at a time, so instead of having multiple single column indexes that can only check one column, a multi-column index was used instead. Indexes are important because they make queries more efficient and run faster. These indexes improve query performance by providing the database the position of the relevant data in the table, making the database jump ahead to that position without having to scan all the other rows in the table while looking for relevant data. This will allow faster querying for the ngram, lng_src, n_number, and date columns since MySQL will know the locations of the data inside the table.

The ngram table contains an auto incremental ngram_id as the primary key, ngram as the ngram, and n_ number being the size of the n-gram (1, 2, or 3). None of these fields can be nulled as they are all used for RegEx querying. The ngram table is indexed on n_number to speed up the queries.

The details (title, author, date, and link) of all the scraped articles are also stored in another table. This table contains the source_id as the auto incremental primary key. The link must be present because this is used to ensure that the same article isn't scraped again. The category of the source must also be present to let the user know what kind of medium the data is from. The author must be present to let the user know who wrote the piece. Lastly, the date must be present because the source cannot appear in the visualization graphs without it.

Lastly, the monthly statistics of the corpora is stored as well. The ngram_count_per_lang contains an auto incremental ngram_count_id as the primary key, lng_src signifies which language the data came from, ngram_count represents the sum of all ngram counts for that month, and the date serves as the date where the data came from. This table is indexed on the multi-column index of date_lng_src_index, this index utilizes the columns date, and lng_src to speed up the queries.

3.5 Data Visualization

For the visualizations, the researchers used Chart.js, a data visualization library in JavaScript.

Line Charts: Line charts are used to visualize the change of a value over time. In the example shown in Figure 3, the chart shows the size of the English lexicon over time from Michel et al. (2011). Line charts will be used to visualize the word count over time.

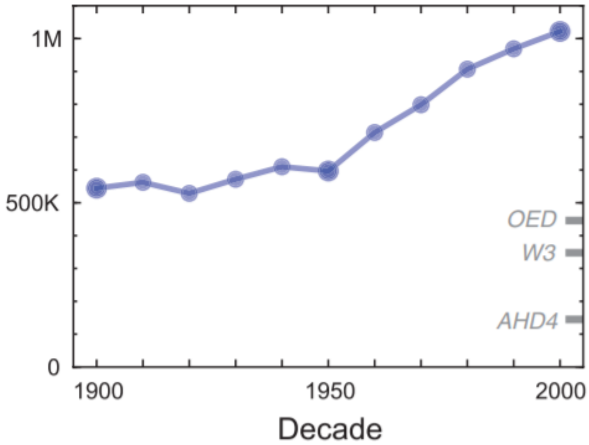


Figure 3. Line chart of the English lexicon and the coverage of the words in three dictionaries. Image taken from Michel et al. (2011).

Frequency Charts: Frequency charts is a variety of a line chart that uses a frequency of a value over time instead of value over time. The chart in Figure 4 shows the usage frequency of the word "slavery" over time, the frequency value is computed by the word count for that year divided by the total amount of words in the corpus for that year (Michel et al., 2011). This chart will be used to find the usage of the word over time, as well as compare it with the usage frequency of other words.

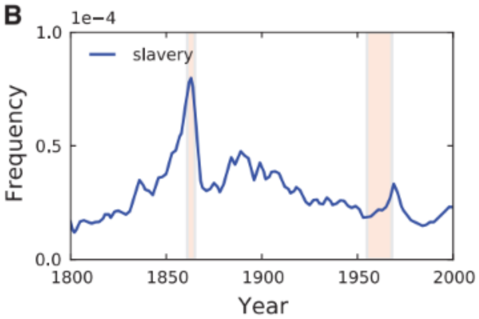


Figure 4. Frequency chart showing the word frequency of the word 'slavery' with red marks highlighting the events of the Civil War and the Civil Rights Movement. Image taken from Michel et al. (2011).

3.6 User Testing

For the user testing, individuals who have backgrounds in linguistics or language analysis were selected, as most of the features in the system are tailored for their use. The testing procedure is split into 4 parts: a background interview of the user, an introduction to the system, a demonstration on how to use the system, and an exit interview.

The background interview consisted of questions regarding their experience in the field, as well as inquiries on how they would normally do analysis on certain words or languages, as well as questions on what technologies or existing systems they use to aid in their analysis.

The system was then introduced to them, and they were given the chance to experiment with the features as well as accomplish certain tasks that are given to them. After the demo phase, a final interview is conducted to gather feedback and suggestions regarding the system.

4 Gunita Features and Interesting Searches

4.1 N-gram Search

This feature allows the user to query for an n-gram from the database and see the n-gram's usage over time through a line chart or a frequency chart.

Pasko vs Christmas: Figure 5 shows the yearly Christmas trend in the Philippines. Instead of having "Christmas" and "Pasko" occur only in December, it can be seen that the Christmas

season starts around September and peaks in December and quickly falling off in January. This graph agrees with the findings of Ilao et al. (2011) regarding the Filipino Christmas season.
The users can interact with the chart to know the exact number of n-gram occurrences in the timeline. When they click on a specific time period in the timeline, they can also see all the sources where the n-gram was mention along with a link to the news source if necessary. This feature provides the user context on why the n-gram gained or lose popularity.

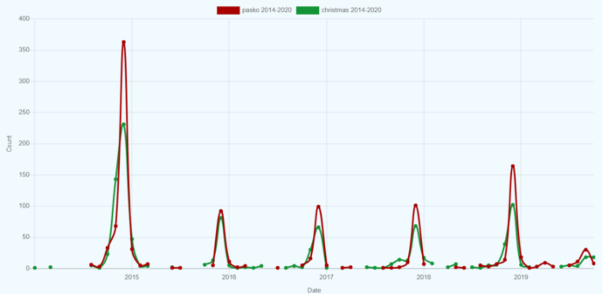


Figure 5. Graph showing the occurrences of "Pasko" (in red) vs "Christmas" (in green) in the collected Philippine news text.

Pilipinas vs Filipinas: Figure 6 shows that "Pilipinas" is more widely used compared to "Filipinas". Despite efforts from the Komisyon sa Wikang Filipino (2013) pushing for the use of the official term "Filipinas", it can be seen that "Pilipinas" is the more preferred version of the word, however the data is primarily from news articles and so, it only shows the more preferred version in news articles. It is also important to note the the recent usage of the term "Filipinas" are present because of news articles reporting on the Komisyon sa Wikang Filipino (2013) proposal to change the Philippine's official Filipino term from "Pilipinas" to "Filipinas".

West Philippine Sea vs South China Sea: Figure 7 shows that the usage of "South China Sea" is generally more prevalent before 2019, while the usage of "West Philippine Sea" is used more after 2019. This could be due to rising political tensions over this region.

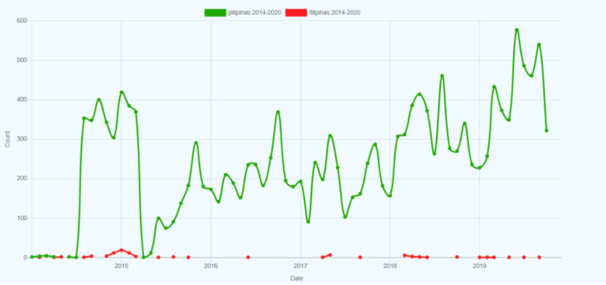


Figure 6. Graph showing the occurrences of "Pilipinas"(in green) vs "Filipinas" (in red). The graph here uses scraped Philippine news text and the text corpus collected by Ilao and Guevara (2012).

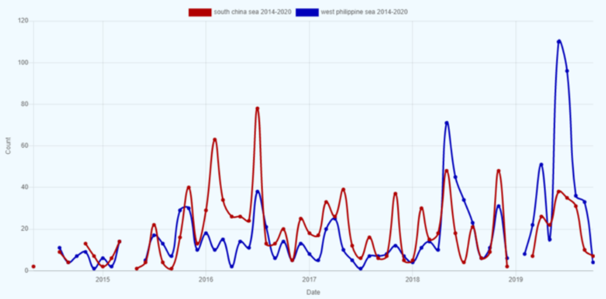
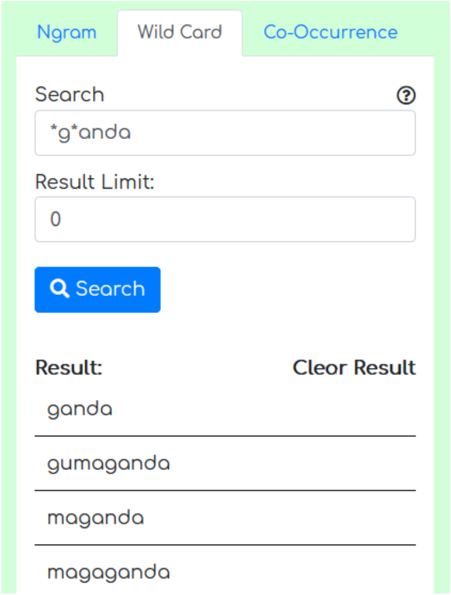


Figure 7. Graph showing the occurrences of "West Philippine Sea" (in blue) vs "South China Sea"(in red) in the collected Philippines new sources.

4.2 Wild card search

The wild card search function allows users to search for the derivations of root words. With the use of a simplified version of regular expressions, users can use a combination of 3 symbols: "*", "+", and "?". The "*" symbol can be used to represent 0 or more characters, the +" symbol can be used to represent 0 or 1 characters, and the "?" can be used to represent any single character.

In the figure below, the word "ganda" is used to serve as an example. The result limit is set to 0 to signify that the user wants to list all n-grams that match the Wild Card search term.



Search
*g*anda

Result Limit:
0

Result: Cleor Result

ganda

gumaganda

maganda

magaganda

Figure 8. Wild card results for "*g*anda" showing the derivations of the word "ganda".

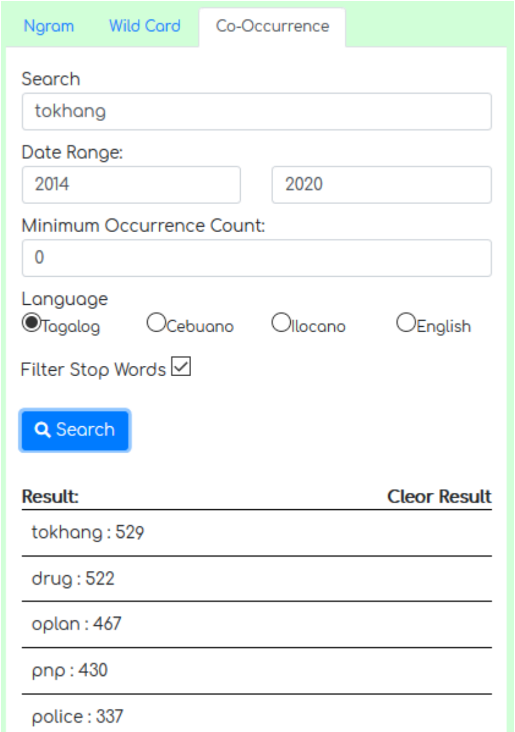
Figure 8 shows that the input *g*anda was able to capture many derivative forms of the root word ganda, although there are some words that aren't derivative forms of "ganda", it is still acceptable. This feature is important since the Filipino language has many variations where pre-, in-, and post-fixes are place in placed in different parts of the derivation.

4.3 Co-occurrence Search

The co-occurrence search allows the user to search for the words that most commonly occur with the given n-gram.

The n-gram "tokhang" will be used to serve as an example. The minimum co-occurrence count can be set to limit the results; it is currently set to 0 to list all co-occurrences regardless of how many times the co-occurrence appeared in the corpus. The user can also select the source language and has the option to filter stop words as well.

The co-occurrence results in Figure 9 show that the word "tokhang" is very commonly used along with the "pnp", the Philippine National Police. While there are many words that aren't related to the word "tokhang" but are instead just extremely common words in the Filipino languages. This search shows that the word "tokhang" is commonly used to refer to the Philippine National Police war on drugs.



Search
tokhang

Date Range:
2014
2020

Minimum Occurrence Count:
0

Language
● Tagalog ○ Cebuano ○ Ilocano ○ English

Filter Stop Words ☑

Result: Cleor Result

tokhang : 529

drug : 522

oplan : 467

pnp : 430

police : 337

Figure 9. The co-occurrence search results for "tokhang" show that "drug", "oplan", "pnp", and "police" are commonly used with the word "tokhang".

5 User Testing

Three experts from the fields of language education, Philippine studiess, and Filipino language tried Gunita and checked what features they can use for their respective research works.

The general feedback from all users is that they had a positive experience with the system since they liked the majority of the features that were shown to them. The most common feedback was that the visualizations were nice and was also very insightful as it highlighted the increase and decrease of the usage of a certain word over time. They also liked the feature of viewing the sources of the points in the graph, as it provided the cause for the increase or decrease as well as a way for the user to look deeper as to why the increase or decrease happened.

Although the users were initially surprised at the large amount of data that Gunita contains, a large portion of it was still lacking especially on dates before 2014 and the scarcity became more evident as the users searched for more specific n-grams. The users also suggested that the system should cover more categories of textual data such as short stories and blog posts and should not be limited only to news articles. They mentioned that including these resources does not only provide more data, but also gives more forms of writing since news articles normally use a formal way of writing as opposed to textual data from social media sites and blog stories. The users also suggested a Parts Of Speech (POS) Tagger on the n-grams to get rid of specific parts of speeches in the co-occurrence results, as well as provide a more specific visualization to an n-gram depending on the context of how it was used. The usage of the RegEx search also confused the users as it required them to know how to use RegEx patterns.

The users all agreed that the system can provide a lot of research potential, and that it would help them in their own research.

6 Conclusion and Future Works

The general objective of visualizing the evolution of written texts in the Philippines was achieved by developing a system that generates the said visualization given certain search parameters. The requirements of the system were determined by reviewing related systems that currently exist, Gunita was then tested by experts to verify the requirements as well as gather feedback on what features could be added or improved upon on the system. After reviewing multiple systems and finishing the development of our three main features (n-gram search, wild card search, co-occurrence search) the general consensus among the experts interviewed was that the system had many applicable uses in the field of linguistic research.

However, there is much room for improvement. The wild card search can be quite limited due to the simple implementation of RegEx. A balance between user accessibility for users unfamiliar with RegEx and thorough searching with complex RegEx must be found.

The system is also extremely biased to recent news articles from 2014-2019, and is not very accurate for mediums of literature outside of news articles since the tone of news articles are formal. For example, trying to visualize and gain insights on particular slang and informal words may not work due to data sparsity. The system would benefit from more sources of data such as WattPad and Twitter.

The system could implement Parts Of Speech (POS) tagging in the future as this would provide better context when visualizing the usage of some n-grams as well as provide better results when searching for related n-grams with Wild Card or Co-occurrence search.

The languages can also be labelled better. In its current state, the system labels whole articles according to its predominant language, so if a Filipino article quotes an English sentence, that English sentence will be classified as Filipino. A solution to this problem would be to implement sentence-level language labelling, this can be done by splitting the contents of the article on the punctuation marks, and then labelling the language per sentence instead of per article.

Acknowledgments

The authors want to thank Dr. Joel Ilao for their generosity in sharing their historical Filipino corpus, and Dr. Charibeth Cheng for their valuable insights on this project. The authors also acknowledge Dr. Shirley Dita, Dr. Rhoderick Nuncio, and Dr. Raquel Sison-Buban for lending their expertise to the researchers on how to improve Gunita.

References

Komisyon sa Wikang Filipino. (2013). Pinagtitibay ang kapasiyahan ng pagbabalik ng gamit “filipinas” habang pinipig ang paggamit ng “pilipinas”.

Michel, J.B., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., & others (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176–182.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday, 16(9).

Ilao, J., Guevara, R., Llenaresas, V., Narvaez, E., & Peregrino, J. (2011). Bantay-Wika: towards a better understanding of the dynamics of Filipino culture and linguistic change. In Proceedings of the 9th Workshop on Asian Language Resources (pp. 10–17).

Ilao, J., & Guevara, R. (2012). Investigating spelling variants and conventionalization rates in the Philippine national language’s system of orthography using a Philippine historical text corpus. Proc. of O-COCOSDA.

Barker, C. (2003). Cultural studies: Theory and practice. Sage.

Multi-Word Verbs in Senior High School Academic Papers: A Corpus-Based Study

Ma. Angelica A. Gumangan
De La Salle University
ma_angelica_gumangan@dlsu.edu.ph

Shirley N. Dita
De La Salle University
shirley.dita@dlsu.edu.ph

Abstract

A multi-word verb refers to a verb-particle-preposition combination which acts a single lexical and syntactic unit (Quirk et al., 1985). Graduate students from a private university compiled a learner corpus composed of academic papers of senior high school students. Using this 1.5-million-word learner corpus, this study looks into the frequency of the most used MWVs and the innovations in the use thereof. According to Quirk et al., phrasal verbs (find out, come along), prepositional verbs (look at, refer to), and phrasal-prepositional verbs (get back to, keep up with). Using the list curated by Ella (2019), this study investigated 66 MWVs in their finite and non-finite forms using AntConc 3.5. Results indicate that 57 MWVs were existent in the corpus. Also, 711 phrasal verbs (finite- 383 and non-finite- 328), 3836 prepositional verbs (finite- 2628 and non-finite-1208), and 194 phrasal-prepositional verbs (finite-116 and non-finite- 78) were discovered. The ten most frequently used MWVs are *base on, refer to, serve as, contribute to, deal with, find out, look for, give NP to, make up, and come up with.* As for the innovations, *base from, base in, come on, come up to, deal with, make up, make up to, result to, set up, and spend in* were present in the learner corpus. To sum, Filipino students are familiar with prepositional and phrasal verbs, but they avoid phrasal-prepositional verbs due to the complexity of the combination.

1 Introduction

Multi-Word Verbs (MWVs from this point onwards) are considered to be one of the toughest lexical items to learn and master among ESL and EFL students (Siyanova & Schmitt, 2007). They seem problematic because of their complicated nature. If learners are not cognizant of their form, then they would have a difficult time deciphering their meaning and function since most of the items do not have a translation or equivalent in their first language (Ella & Dita, 2017).

According to Quirk et al.(1985), MWVs are structures that behave as a single unit. Examples are deal with, carry out, look forward to. They are a combination of verbs and are more present in conversations than in written form (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Siyanova & Schmitt, 2007).

Perhaps the simplest definition of MWVs is that they consist of two or more words which work together (Siyanova & Schmitt, 2007). That being said, it would be difficult and challenging to realize the said construction as a single semantic and grammatical unit.

It must be noted that long before Quirk et al. (1985) has laid a solid foundation on MWVS in their book, *A Comprehensive Grammar of the English Language*, several scholars have been researching on this grammatical item. They have also been using different terminologies for MWVs-phrasal verb (Bolinger, 1971), group verb (Sweet, 1892), compound verb (Hook, 1974), verb-particle combination (Fraser, 1976), separable and inseparable verbs (English Language Services, 1964), and multiword constructions (Moon, 1998), to name a few. However, as mentioned earlier, it was Quirk et al. (1985) who have eloquently established grounded rules on what comprises MWVs and their types.

MWVs serve as a good barometer of competence for they reflect high fluency in the English language (Howarth, 1998; Liu, 2012). It is said that multiword expressions are an integral part that affects language ability (Wang, 2020) since they are extremely widespread in and representative of the English language (Tu & Thao, 2019).

Several articles had been published on the frequency and occurrence of MWVs in academic articles. Biber et. al (1999), Gardner and Davies (2007) and Trebits (2009) gave focus on the most frequent MWVs in different corpora and registers.

Some researchers looked into the phrasal verbs in Malaysian ESL learners' narrative compositions from The English Language of Malaysian School Students or EMAS corpus (Akbari, 2009). Similarly, Mendis (2010) examined the use/non-use of phrasal verbs in two varieties of English- British English and Sri Lankan English. Chen (2013) used a 188,628-word learner corpus and explored the overuse or underuse of phrasal verbs. Adding to this, Ryoo (2013) investigated the use of phrasal verbs in student essays. Kartal (2008) determined the phrasal verbs in four Turkish university-level coursebooks and pointed out the most frequent phrasal verbs in the COCA and BNC. Ngoc and Thao (2019) investigated the use of phrasal verbs in the research proposals among 19 Vietnamese graduate students. Azmin Md Zamin et al. (2019) studied the use of phrasal verbs in Malaysian secondary school textbooks. Interestingly, Ekasani, Yadnya, Artawa, and Indrawati (2019) conducted a study in which they looked at the MWVs found in an English cookbook translated into Indonesian. Alangari, Jaworska, and Laws (2020) investigated the frequency and meanings of phrasal verbs in expert academic writing in the discipline of linguistics. Divišová (2020 aimed to give quantitative and qualitative analysis of MWVs in the speech of native and nonnative speakers of English.

Results and findings of these studies had led to the creation of word lists such as A Pedagogical List of Phrasal Verbs (PhaVE List) by Garnier and Schmitt (2015). Furthermore, phrasal verbs dictionaries (i.e., *Oxford Phrasal Verbs, Longman Dictionary of Phrasal Verbs, American Heritage Dictionary of Phrasal Verbs, and Cambridge International Dictionary of Phrasal Verbs*) were created and produced to aid linguists and language teachers in understanding this grammatical item. Most importantly, the studies mentioned above inspired more researchers and language teachers to know more about MWVs and their nature.

In the local setting, only four studies were known to scholars of MWVs. Bensal (2012) studied the prepositional verbs present in Asian Englishes. Ella and Dita (2017) described the phrasal-prepositional verbs in the Philippine component of the International Corpus of English. Ella (2019), in her dissertation, explored the MWVs reflected in 12 World Englishes using the International Corpus of English. Recently, Somoson (2020) looked into the phrasal verbs found in Philippine English using ICE-PH. The research mentioned previously all used corpora from written and spoken registers.

As can be seen, there is an evident scarcity in the literature of MWVs and its types in the local setting. The lack of available studies may result to incognizance of researchers and language teachers on such grammatical forms. Thus, it is the intent of this paper to add to the body of literature on MWVs. Consequently, the following questions were answered in this paper:

1. What is the frequency of multi-word verbs in the learner corpus?
2. What innovations in the use of MWVs were found in the learner corpus?

1.1 Theoretical Framework

Multi-word verbs consist of two or more words which work as together (Siyanova & Schmitt, 2007). Quirk et al., (1985) stated that MWVs are *chucks of syntax* made up of a verb proper and a particle. The meaning of MWVs cannot be guessed or inferred by simply looking at each lexical item. Thus, MWVs must be looked at as a single lexical and syntactic unit. There are three kinds of MWVs in the English language- phrasal verb, prepositional verb, and phrasal-prepositional verb.

Phrasal verbs follow a verb-particle combination. The particle depends on the verb for it to be meaningful. These particles are mostly prepositions (against, among, beside, into, upon, with) or spatial adverbs (ahead, away, back, out). However, there are particles that can be used as a preposition or an adverb (above, across, through under). Examples of phrasal verbs are *turn on, hand in, fond out, blow up, get by, put away, switch on,* and *pick up.*

A prepositional verb is formed by putting together a verb and a preposition that is closely associated or related to it semantically or syntactically (Quirk et al., 1985). In order to distinguish a prepositional verb, De Haan (1988) suggests identifying whether the object of the preposition answers *who* or *what.* If it does, this verb combination functions as a prepositional verb. Common examples of prepositional verbs are *drive around, look after,* and *call on.*

Lastly, phrasal-prepositional verb is a combination of three elements- lexical verb, adverb, and preposition. Claridge (2020) described this verb combination as the "offspring" of phrasal verbs and prepositional verbs. This verb combination, according to Quirk et al. (1985), are mostly used in informal contexts. Examples of phrasal-prepositional verbs are *keep up with, get out of, hold on to,* and *keep away from*.

2 Methodology

A corpus of 1.5 million words was used in the conduct of the study. The corpus consists of academic papers by senior high school students in the Philippines. The corpus, which was compiled by graduate students from a private university in the Philippines as part of their final requirement, is made up of 233 academic papers or 798 .txt files from the different strands in senior high school (See Table 1). However, this learner corpus is not yet available to the public and only those who collated the corpus have access.

Track	Number of Research Papers	Number of .txt files	Number of Words
Accountancy and Business Management (ABM)	40	156	286793
Arts and Design (ADT)	2	6	12589
Computer and Communications Technology (CCT)	2	12	22498
Culinary Arts (CAR)	3	11	18889
General Academic Strand (GAS)	20	48	85306
Humanities and Social Sciences (HUMMS)	55	193	368556
IT in Mobile Application (ITM)	11	46	88075
Science, Technology, Engineering, and Math (STEM)	77	265	517852
Technological and Vocational Livelihood (TVL)	23	61	110957
Total	**233**	**798**	**1511515**

Table 1. Breakdown of the learner corpus

During the finalization of the corpus, it was converted into .txt files, in which unnecessary parts (i.e., title page, table of contents, tables, charts, references, appendix) of the academic papers were omitted. Only the introduction, methodology, results, discussion, conclusion, and recommendation were included in the .txt files.

The AntConc 3.5 software was utilized in this study. This software, which is created by Lawrence Anthony, is mostly used by researchers who study corpus linguistics in finding out the frequency and collocations of words. This software has seven programs to choose from: Concordance Tool, Concordance Plot Tool, File View Tool, Clusters/N-Grams, Collocates, Word List, and Keyword List. For the interest of this study, the KWIC (KeyWord In Context) was specifically used.

The study was guided by the list of MWVs (See Table 2) that was adapted from Quirk et al. (1985) and Ella (2019). There are twenty-two items for each kind of multi-word verb- phrasal verb (PV), prepositional verb (PrV), and phrasal-prepositional verbs (PPV), as seen in the table below. These MWVs were searched in their finite (past and non-past) and non-finite forms (infinitive, gerund, and participles).

Phrasal Verbs	Prepositional Verbs	Phrasal Prepositional Verbs
carry out	apply to	set up in
come along	base on	catch up with
come on	contribute to	come in for
come over	contribute with	come out of
end up	cope with	end up with
find out	deal with	get away from
get in	derive from	get away with
get off	fill NP with	come up with
get out	give NP to	get back to
get up	go through	get off at
give up	look at	get off with
go ahead	look for	get on with
go off	make of	get out of
live in	obtain NP from	go along with
look up	pay for	go out for
make up	refer to	go over to
pick up	rely on	go up to
put on	send NP to	hand over to
set up	serve as	hold on to
stand up	stare at	keep up with
take on	turn to	look out for
take over	use NP as	turn away from
take up	wait for	turn back to

Table 2. List of Multi-Word Verbs

To ensure that the criteria and description of multi-word verbs were met, manual weeding was done. After each search, the researcher looked at the verb combinations to make sure that they functioned as verbs, and not as nouns or adjectives.

3 Results

3.1 Frequency of Phrasal Verbs in Academic Papers of Senior High Students in the Philippines

Among the 22 phrasal verbs, 20 were found in the learner corpus. There were 328 instances of non-finite phrasal verbs and 383 finite phrasal verbs. For finite phrasal verbs, the five commonly used in the corpus are *find out, make up, take up, end up,* and *carry out.* Meanwhile, the five commonly used non-finite phrasal verbs are *find out, carry out, give up, take up,* and *make up.*

Table 3 presents the overall frequency of phrasal verbs in the learner corpus. *Find out, make up, carry out, give up,* and *take up* are the commonly used phrasal verbs in the corpus.

Rank	Phrasal Verb	Frequency
1	find out	242
2	make up	138
3	carry out	76
4	give up	54
5	take up	50
6	end up	29
7	live in	21
8	set up	19
9	stand up	16
10	take on	12

Table 3. Overall frequency of phrasal verbs

As for the least used phrasal verb in the learner corpus, *put on* only has two occurrences in. Similarly, *look up, come along, go off, come on, go ahead, come over,* and *get off* have no occurrences at all.

3.2 Frequency of Prepositional Verbs in Academic Papers of Senior High Students in the Philippines

Twenty-two prepositional verbs were searched, and it was found out that 18 were present in the corpus. There were 2628 occurrences of finite prepositional verbs while there were 1208 occurrences of non-finite prepositional verbs. *Base on, refer to, serve as, contribute to,* and *deal with* are the five commonly used finite prepositional verbs whereas *base on, deal with, look for,* and *contribute to* whereas *base on, deal with, look for, contribute to,* and *give NP to* are the five commonly used non-finite prepositional verbs.

Rank	Prepositional Verb	Frequency
1	base on	1300
2	refer to	482
3	serve as	343
4	contribute to	290
5	deal with	258
6	look for	155
7	give NP to	139
8	look at	105
9	rely on	101
10	apply to	98

Table 4. Overall frequency of prepositional verbs

It can be seen from the table above that *base on, refer to, serve as, contribute to,* and *deal with* are the commonly used prepositional verbs by students in their academic papers. On the other hand, *fill NP with* had the least occurrence in the corpus, with only 12 hits. Similarly, *contribute with* and *stare at* had no occurrence in the corpus.

3.3 Frequency of Phrasal-Prepositional Verbs in Academic Papers of Senior High Students in the Philippines

Fifteen out of 22 phrasal-prepositional verbs were found in the corpus. There were 116 occurrences of finite phrasal-prepositional verbs while 78 non-finite phrasal-prepositional verbs were evident in the learner corpus. *Come up with, make up of, get out of, end up with,* and *go along with* are the five most used finite phrasal-prepositional verbs in the learner corpus while *come up with, keep up with, hold on to, catch up with,* and *come out of* are the most evident non-finite phrasal-prepositional verbs in the academic papers of senior high school students.

Rank	Phrasal Verb	Frequency
1	come up with	114
2	make up of	20
3	keep up with	12
4	get out of	8
5	end up with	7
6	catch up with	6
7	hold on to	6
8	go along with	5
9	come out of	4
9	get back to	4

Table 5. Overall frequency of phrasal-prepositional verbs

As indicated in the table above, come *up with, make up of, keep up with, get out of,* and *end up with* are the most common phrasal-prepositional verbs found in the corpus. Conversely, get away *from, get on with, go out for,* and *go up to* only had one occurrence each in the academic papers. On a similar note, *come in for, get away with, get off with, go over to, look out for, turn away from,* and *turn back to* did not appear in the corpus.

3.4 Overall Frequency of Multi-Word Verbs in Academic Papers of Senior High Students in the Philippines

The finite (-ed,e/es form) and non-finite (-ed/en, -ing, to) forms of the sixty-six multi-word verbs were searched in the corpus. Of the MWVs on the list, 59 MWVs were evident in the learner corpus. The ten most common MWVs, as seen on Table 6, are *base on, refer to, serve as, contribute to, deal with, find out, look for, give NP to, make up,* and *come up with*. It can be gleaned from the table that out of the top ten multi-word verbs, seven were prepositional verbs, two were phrasal verbs, and only one was phrasal-prepositional verb.

On the other hand, the MWVs with the least number of occurrences in the learner corpus are *get away from, get on with, go out for, go up to, come on*, and *go ahea*d, with just one occurrence in the entire corpus. Similarly, come *in for, get away with, get off with, go over to, look out for, turn away from, turn back to, contribute with, stare at, come over,* and *get off* have no occurrence in the corpus.

Rank	Multi-word Verb	Frequency
1	base on	1300
2	refer to	482
3	serve as	343
4	contribute to	290
5	deal with	258
6	find out	242
7	look for	155
8	give NP to	139
9	make up	138
10	come up with	114

Table 6. Top 10 Multi-Word Verbs in the Learner Corpus

The results above confirm the findings of Biber (1999), Zareva (2016), Theyerl (2018), Ella (2019), and Divišová (2020) - prepositional verbs are more common to users of the English language. Also, they reiterated that phrasal-prepositional verbs had the lowest frequency among the three categories. As for spoken corpus, Zareva noted (2016) that prepositional verbs are "twice more prominent than" phrasal verbs and phrasal-prepositional verbs.

3.5 Innovations in the use of MWVs in the learner corpus

The present study also sought to look for the innovations in the use of phrasal verbs, prepositional verbs, and phrasal-prepositional verbs. This was done by carefully examining the entries for each search.

In a study done by Wilcoxon (2014), she used the terms "misuse" and "incorrect usage" in identifying prepositional verbs that do not conform to the standard or prescribed form of the grammatical item.

However, for in the interest of this study, the researchers decided to use the word "innovation" in addressing non-standard forms of MWVs found in the learner corpus. As Borlongan (2017) states, "Philippine English is developmentally progressing"; thus, the avoidance of the terms "misused" and "incorrect usage".

Table 7 shows a list of innovations in the use of MWVs as reflected in the learner corpora made up of senior high school academic papers.

Innovation	Frequency	Prescribed Word
result to	187	result in
base from	115	base on
cope up with	91	cope with
spend in	68	spend on
base in	13	base on
come up to	10	come up with
make up to	7	make up with
come on	4	come to

Table 7. Innovations in the use of MWVs

The sentences below show how these innovations in the use of MWVs were reflected in the learner corpus, comprised of senior high school academic papers.

1. The survey will *result to* the factors that affect the non-Metro Manila migrants in Metro Manila residence. <GAS_04_04>

2. The test is a modified set of questions *based from* a previous study on assessing household preparedness for earthquakes. <STEM_18_03>

3. They are even great therapists that help patients recover and *cope up with* illnesses such as mental diseases and disorders because they give non-judgemental (sic) environments to the sick people. <HUMMS_06_02>

4. He stressed that virtues are the most valuable possessions of human beings and life should be *spent in* the search of goodness. <STEM_06_01>

5. A path model *based in* a theory of social capital was tested with Latino middle school and high school students. Most participants

were immigrants (predominantly from Mexico). <ITM_08_02>

6. The researcher will *come up to* an action plan specifically by means of proposing a project which is implementation of seminar (Different management style and marketing style) for the business owners. <ABM_09_02>

7. They are some of the students reported have a problem when it *comes on* Mathematics and problem solving. <HUMMS_44_02>

The researchers also investigated the occurrences of the enumerated innovations in other corpora, like News on the Web (NOW) Corpus, Global Web-Based English (GloWbE) Corpus, and International Corpus of English- Philippines (ICE-PHI) for comparison.

It can be gleaned from the table that *result to* (1) has more occurrences in the corpus, which takes the place of result in, the prescribed form. Upon searching another corpus, NEWS on the Web (NOW) Corpus, it was revealed that among other varieties of English, Philippine English has the second most occurrence of result to, having 1539 hits.

Base from (2) has 115 occurrences. The prescribed form is *based on*, which means *to use ideas or facts to develop something.*

Another innovation seen in the corpus is *cope up with (3)*, with 91 hits. Its prescribed form is cope with, which means to *survive or deal with*. In the corpus of Global Web-Based English (GloWBE), it was found out that this verb combination is not only present in Philippine English, but also in other Englishes. Indian English had the highest occurrence of cope up with (f=133). This is followed by Bangladeshi English (f=46), Pakistani English (f=40), and Philippine English (f=24). In the NOW Corpus, there were 2711 hits on *cope up with* for over ten years.

Spend in (4), with 68 hits, was also an innovation in the use of MVS. Taking the place of spend on, it has the meaning of *"to expend some amount of time doing or working on*". There were 3102 occurrences of *spend in* across Englishes in the Global Web-Based English (GloWBE). A search on the NOW Corpus revealed 20769 hits on this innovation.

Base in (5) has 13 hits from the corpus, and has the same meaning as base on. However, there are instances in the corpus in which *base in* (8 & 9) was used correctly and which means *to operate from a particular place or location.* Examples are given below:

8. Jollibee is a Filipino multinational chain of fast-food restaurants *based in* Pasig, Philippines. <TVL_05_01>
9. Defend 2514, a Land Enraged, is a role-playing game by Anino Entertainment, an independent video game company *based in* the Manila, Philippines (Velvey, 2003). <ITM_10_02>

Come up to (6), with 10 instances, was used in the corpus to mean *produce something,* which is the definition of come up with, the prescribed form of this phrasal-prepositional verb.

Lastly, *come on* (7), which means "*to reach or be brought to a specified situation or result"* had 4 occurrences in the corpus. Its prescribed form is come to.

4 Discussion

Looking at the results of the study, it shows that senior high school students are already familiar with the different verb-adverb-preposition combination in the English language. This claim is evident in the frequency of phrasal verbs, prepositional verbs, and phrasal-prepositional verbs found in the learner corpus.

The results of the study show that prepositional verbs are more common and prevalent in the academic papers, representing 80.91% of the MWVs and .25% of the whole corpus. Phrasal verbs comprise 15% of the sum of the MWVs found in the study, which is about .05% of the learner corpus. Lastly, phrasal-prepositional verbs only has 4.09% presence in the totality of MWVs and about .01& of the whole corpus. These numbers are congruent with the results of Biber et al. (1999), Zareva (2006), Ella (2019) and Divišová (2020).

It is certain to conclude that Filipino learners are already familiar with the verb + preposition combination since prepositional verbs appeared more in the learner corpus compared to the other verb combinations. *Base on*, with 1300 counts in the corpus, was mostly used in the form d/ed, both in past and participial form. It must also be noted that *base from* and *base in,* an innovation of *base on*, had gained 115 and 13 instances, respectively. These innovations have the same definition as the standard *base on*.

The results show that prepositional verbs were mostly used in their finite form, specifically in the past tense. As for the non-finite form, this verb combination was mostly used as past participle form. Among the ten most used prepositional verbs in the finite and non-finite form, *base on, deal with, look for, contribute to, give NP to,* and *rely on* were present in both forms. It can be noted that the finite form of the prepositional verbs has a higher occurrence and mean than the non-finite form. This implies that students are more familiar with the finite form of prepositional verbs more.

The results of the present study substantiate the findings of Seilhamer (2003), Liu (2011), Wilcoxon (2014), and Ella (2019). Most of the top ten prepositional verbs in the present study were also present in the studies cited earlier.

Find out, make up, carry out, give up, take up, live in, stand up, and *take on* were eight of the top ten overall phrasal verbs that were common in both finite and non-finite forms. The base form of the finite phrasal verbs was mostly used while for the non-finite form, the infinitive to was primarily evident. The results of this study show that the top ten most used phrasal verbs in senior high school academic papers in the Philippines were likewise discovered in studies by Garner and Davies (2007), Liu (2011), Garnier and Schmitt (2015), and Ella (2019). *Make up, carry out, and take up* were also found in the studies mentioned earlier. This means that these verb combinations are evident in the present study and in the corpora used by other researchers.

As for phrasal-prepositional verbs, this verb combination has by far the least occurrence in the corpus. This conclusion was also pointed out by Ella and Dita (2017). They mentioned in their study that Filipinos reflect slight use of phrasal-prepositional verbs in the corpus. This verb combination was mostly used its finite past tense and non-finite infinitive to form. In comparison to phrasal verbs and prepositional verbs, the outcomes of this study imply that students are unfamiliar with phrasal-prepositional verbs. It is possible that they have not encountered the verb + preposition + particle combination before, resulting in the form's unfamiliarity.

It is evident that learners are still moderate and passive in the use of this verb-adverb-preposition combination. One of the reasons for this phenomenon is that MWVs have no equivalent grammatical item in Filipino. *Get out of*, *catch up with,* and *hold on to* are three of the phrasal-prepositional verbs present in this study, Ella and Dita (2017) and Ella (2019). Another reason for this occurrence might be due to the lack of direct instruction on this grammatical item.

As ascertained in the results, there is evidence that there are innovations in the use of MWVs in the written discourse. Interestingly, some of the innovations discovered in this study were also present in other varieties of English, such as in the case of *result in* (result in), *cope up with* (cope with), and *base from* (base on).

Limited exposure to the different verb-adverb-preposition might play a role in the innovations in the use of multi-word verbs in written and spoken discourse. Confusion on which adverb and preposition to use might also be a factor for such occurrence (Ella & Dita, 2017).

Prepositions are difficult to learn, according to Seilhamer (2003), even for native speakers. Lindstromberg (2001) pointed out that only about 10% of EFL students can use and understand prepositions. Prepositions are difficult to learn and grasp, according to Seilhamer (2003), because of how they differ across languages.

Let us look at the case of Japanese speakers. Yasuda (2010) observed that Japanese students lack the cognizance of particles and their meanings. As a result, they have conflicts in using particles to their references.

The French also do not have verb and particle constructions in their mother tongue, which then causes confusion among them in using phrasal verbs.

Bautista (2000) discovered two notable findings for Philippine English. *Based from* and *result to* seem to be frequent among users of English in the Philippines. She explained that prepositions in Filipino seem to have an all-purpose form and function- *sa* and *ng.*

Salazar (2022) mentioned that Filipinos have the tendency to use *cope up with* instead of cope with for the former follows the phrasal-prepositional verb *keep up with.*

5 Conclusion

The study sought to investigate the frequency of multi-word verbs in their finite and non-finite forms using a corpus from academic papers of senior high school students in the Philippines. Also, the researchers wanted to look for innovations in the use of MWVs in the corpus. There were 4741 counts of MWVs from the 1.5-million-word corpus. There were 3836 prepositional verbs, 711 phrasal verbs, 191 were phrasal-prepositional verbs. Most of the MWVs were used in the finite form, in the past tense. While for the non-finite form, the past participle was used.

Of the 66 MWVs searched in the corpus, 59 were evident. The ten most used MWVs are *base on, refer to, serve as, contribute to, deal with, find out, look for, give NP to, make up* and *come up with*. Of the ten MWVs, there were seven prepositional verbs, two phrasal verbs, and one phrasal-prepositional verbs. This distribution of MWVs are consistent with the study of Biber et al (1999), Zareva (2006), Ella (2019) and Divišová (2020). Learners are more familiar with prepositional verbs as reflected in the results; however, they are cautious in using phrasal-prepositional verbs. Limited exposure to this verb combination and unfamiliarity might be factors.

There were also instances of innovations in the use of MWVs in the corpus. These are *result to, base from, cope up with, spend in, base in, come up to, make up to,* and *come on*. It is also notable that these innovations are not only found in Philippine English, but also in other varieties of English.

Keeping the results in mind, Filipino senior high school students are already familiar with the concept of multi-word verbs as evident in their academic papers.

As for the pedagogical implications of the study, it is suggested that the perceptions, perspectives, and preferences of language teachers be looked into to strengthen the teaching and learning process of multi-word verbs. It would also be interesting to comprehend how language teachers view this MWVs. Furthermore, a more in-depth study on MWVs can be accomplished to produce learning materials to help Filipino learners become more cognizant of phrasal verbs, prepositional verbs, and phrasal-prepositional verbs.

References

Ainul Azmin Md Zamin, Mahmoud Elfeky, Rafidah Kamarudin, Faizah Abd Majid. 2019. A corpus-based study on the use of phrasal verbs in Malaysian secondary school textbooks. International Journal of Applied Linguistics and English Literature, 8(6), 76- 85. https://doi.org/10.7575/aiac.ijalel.v.8n.6p.76

Anna Siyanova and Norbert Schmitt. 2007. Native and nonnative use of multi-word vs. one-word verbs. IRAL - International Review of Applied Linguistics in Language Teaching, 45(2), 119-139. https://doi.org/10.1515/iral.2007.005

Claudia Claridge. 2000. Multi-word verbs in Early Modern English: A corpus-based study (No. 32). Rodopi. Dandie Somoson. (2020). A Corpus-Linguistic Analysis of Phrasal Verbs in Philippine English. In 1st International Conference on Information Technology and Education (ICITE 2020) (pp. 135-140). Atlantis Press.

Danica Salazar. 2022. Introduction to Philippine English. Oxford English Dictionary. https://public.oed.com/blog/introduction-to-philippine-english/

Dee Gardner and Mark Davies. 2007. Pointing out frequent phrasal verbs: A corpus-based analysis. TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect, 41 (2), 339-359.

Douglas Biber, Geoffrey Leech, Stig Johansson, Susan Conrad, and Edward Finegan. 1999. Longman grammar of spoken and written English. Pearson Education Limited, Harlow, England.

Dushyanthi Mendis. 2010. Formality in academic writing: The use/non-use of phrasal verbs in two varieties of English. In M. Ruiz-Garrido, J. Palmer-Soliven, I. Fortanet-Gomez (Ed.), English for Professional and Academic Purposes (pp.9-23). Rodopi.

Edwina Bensal. 2012. The prepositional verbs in Asian Englishes: A corpus-based analysis. Unpublished Master's Thesis, De La Salle University Manila.

Galip Kartal. 2018. A corpus-based analysis of phrasal verbs in ELT coursebooks used in Turkey. Cumhuriyet International Journal of Education, 7(4), 534-550. https://doi.org/10.30703/cije.466035

Jennibelle Ella. 2019. Multi-word verbs across Englishes: a corpus-based study. [Unpublished Dissertation]. De La Salle University.

Jennibelle Ella and Shirley Dita. 2017. The Phrasal-Prepositional Verbs in Philippine English: A Corpus-based Analysis. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (pp. 34-41). University of the Philippines, Cebu.

Kadek Ayu Ekasani, Ida Bagus Putra Yadnya, Ketut Artawa, and Ni Luh Ketut Mas Indrawati. 2019. Translation of multi-word verbs in English cookbook into Indonesian. International Linguistics Research, 2(2), 36-41. https://doi.org/10.30560/ilr.v2n2p36

Klara Divišová. 2020. Multi-word verbs in speech of native and non-native speakers of English [Master's Thesis]. Univerzita Karlova. https://dspace.cuni.cz/handle/20.500.11956/118546

Manal Alangari, Sylvia Jaworska, Jacqueline Laws. 2020. Who's afraid of phrasal verbs? The use of phrasal verbs in expert academic writing in the discipline of linguistics. Journal of English for Academic Purposes, 43 (1), 1-13. https://doi.org/10.1016/j.jeap.2019.100814

Maria Lourdes Bautista. 2000. Studies of Philippine English in the Philippines. Philippine Journal of Linguistics, 31(1), 39-65.

Meilin Chen. 2013. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. International Journal of Corpus Linguistics, 18(3), 418-442. https://doi.org/10.1075/ijcl.18.3.07che

Omid Akbari. 2009. A corpus-based study on Malaysian ESL learners' use of phrasal verbs in narrative compositions [Doctoral dissertation]. Universiti Putra Malaysia.

Pieter de Haan. 1989. Postmodifying clauses in the English noun phrase: A corpus-based study (No. 3). Rodopi.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985.A comprehensive grammar of the English language. Longman, London.

Tran Phan Ngoc Tu and Tran Quoc Thao. 2019. The use of phrasal verbs in English language research proposals by Vietnamese MA students. VNU Journal of Foreign Studies, 35(4), 114-129. https://doi.org/10.25073/2525-2445/vnufs.4399.

Early writing of a bilingual child: A content analysis of his YouTube video comments

Bernard M. Barruga
Dr. Emilio B. Espinosa, Sr. Memorial State College of Agriculture and Technology/
Mandaon, Philippines
Ateneo de Naga University/ Naga, Philippines
bbarruga@gbox.adnu.edu.ph

Abstract

Language and literacy development has come to be influenced by digital technology in the current information age. Comments posted on a video streaming site by a developing bilingual seven-year-old child speaking an autochthonous Philippine language and English were analyzed as to their mean length, syntactic form and pragmatic uses. Results revealed a language development in early English writing comparable to native speaker oral norms extant in the literature. Implications for further language and literacy development of the participant especially in the use of internet and digital technology in learning as well as for language policy in education are also discussed.

Keywords: mean length of utterance, multilingual education, child language, YouTube, digital literacy

1 Introduction

Contemporary society has placed an outsized role for language and literacy skills of learners. Early childhood programs have been touted to help in this regard. As evidence, the full development of young children focusing on early childhood programs has attracted the attention of scholars and development agencies worldwide (Engle, et al., 2007). Naturally, parents of children in the emergent literacy stage also place a premium on the latter's development of literacy skills.

The current situation in homes with both parents working leaves young children with their home child minders and increasingly with internet sites like YouTube. Early literate children who interact through online gaming, social media and video sites may offer a glimpse into their language and literacy development. Thus, discourse is no longer limited to in-person interactions but also technology-mediated platforms (Waring, 2018, p. 7).

Children developing their language skills must not just master the systems of sounds, meanings, and word order but must learn communicative competence, too (Bryant, 2009). Communicative competence, also known as pragmatic skills, are needed for effective interactions with other people such as peers, families, and teachers. It is also predictive of later academic achievement (Reeder, Shapiro, Watson, and Goelman, 1996). e pages.

1.1 Theoretical framework

Sociocultural theory (SCT): This theory posits that human psychology comes as a result of "a trichotomy of *social practice* (behavior), *consciousness* (cognition and affect), and *material culture* (artifacts, such as language)" (van Compernolle, 2022, citing Blunden, 2011). In this theorization, the three components mediate one another. For instance, communicative speech or writing like commenting on a viewed YouTube video is mediated by participation in a social practice like watching a gamer stream himself playing a game, both of which are also mediated by the mental and emotional processes of cognition and affect. Sociocultural theory is often linked with Vygotsky's (1978) work on the zone of proximal development (ZPD) to take advantage of extraneous resources in helping a learner learn a target language.

According to Snyder and Tour (2017), the sociocultural conception of literacy became more mainstream in the 1980s, necessitating the use of more ethnographic approaches and thus, in the area of computer-mediated communication, varied viewpoints were employed, aside from studying quantitatively the impact of computers on writing, for instance. Later, in the 1990s, qualitative methods became the preferred mode for exploring the intersection among literacy, learning, and technology, still with computer use at the forefront of the research contexts. Eventually, the internet became a frequent site for research, with the sociocultural perspective becoming a focal lens for comprehending literacy (p. 520).

Significant literature exemplified the attempts to untangle the complex relationship of literacy, learning, and technology (Turkle, 1995; Burbules and Callister, 2000; Takayoshi et al., 1999; Kress and van Leeuwen, 2001; Lankshear and Knobel, 2011; Lankshear, et al., 2000; Warschauer, 2006). Gee (2003), meanwhile, has recognized “nonschool literacies” from computer and video games and their study as something of value for education. The studies of Marsh (2011) and Livingstone (2002, 2012) provide evidence for the role and value of digital technology in learning among children and young people.

Bilingualism and internet media: In the early period of media studies in education, practices at home received little consideration as having legitimate value in learning contexts (de Saint-Georges, 2017). Nevertheless, media became a tool of education within the classroom or for independent learning in the decades from the 1970s to the 1990s and continues to the present. Media became the first steps into literacy for many learners (p. 115). There is, thus, a recognition of an autonomous form of learning in which researchers have advocated for the use of media in language learning (Barton and Lee, 2013). In turn, bilingual and multilingual learners benefited from this democratization of second and foreign language learning.

A certain trend of research emerged which looked at the practices of the users of the media (Kelly-Holmes and Milani, 2013). de Saint-Georges (2017) advocated for the study of the practices of media consumers and producers in addition to the focus on the text and language in communication media (p. 119). At this juncture, sociolinguistic approaches to bilingualism become relevant in which the bilingual or multilingual person is seen as a social actor (Wei, 2017, p.221). In the context of video and computer gaming which has become prevalent as of late, viewers who watch players-cum-video streamers connect with the gaming community in a negotiation of identity. The phenomenon has happened cross-culturally, with mostly native American or Canadian English-speakers producing the videos for an international audience via YouTube and other social media sites.

1.2 Background of the study

The global pandemic resulted in early elementary pupils being unable to attend in-person classes regularly. Parents had to act as teachers at home and impart instruction to their own children. Starting in 2020, my own youngest child, who was five years old then, had to answer printed modules for his kindergarten studies. A year later, as a first grader, it was more of the same. As his parents, my wife, who is also an elementary grades teacher, and I endeavored to at least give him beginning reading instruction. Around the middle of 2021, we were successful in making him read one- to three-syllable English words, both through sight word reading and phonological recoding. In other words, he was able to read through the dual routes (Coltheart, 2005).

1.3 Research questions

This study aimed to analyze the YouTube comments as utterances of a bilingual (Minasbate and English) child as to their syntactic type (simple or complex sentence) and pragmatic purposes. The following are the specific research questions: (a) What is the written language development level of the child measured as mean length of utterance of his YouTube video comments; (b) What types of sentences (simple or complex) did he generate; (c) What communicative purposes did he use; and (d) What are the implications for his further language and literacy development?

1.4 Definition of terms

Mean length of utterance – the number of morphemes in a single utterance (Brown, 1973). Normally, the mean length of utterance is used to measure oral discourse. However, for this study, utterance is taken to be written comments.

Utterance – a single utterance is a single written comment posted

Simple sentence – a sentence with one independent clause (Lust, Foley, and Dye, 2009)

Complex sentence – a sentence that is not a simple sentence (Lust, Foley, and Dye, 2009)

2 Methodology

This study made use of the content analysis technique “to study human behavior in an indirect way, through an analysis of their communications. It is just what its name implies: the analysis of the usually, but not necessarily, written contents of a communication" (Fraenkel, Wallen, and Hyun, 2011, p. 478).

The participant is a seven-year-old first grader undergoing instruction through printed modules, although such modules are irregularly accomplished by him. He is the youngest child in a brood of three by the researcher and his wife. He is often left at home with his older brothers who are senior high school students and usually spends time on a cell phone watching YouTube videos of streamers playing games like Roblox and Minecraft, which he also plays from time to time. He also watches content related to vehicles like trains and cars, and some movie clips. His first language is Minasbate western dialect with a sprinkling of Tagalog words as his previous primary child minder spoke Tagalog. He knows all of the letters of the English alphabet since he was about three years old, learning them through YouTube videos and alphabet songs. For material his suitable to his age, he can also speak and comprehend oral English, and since about the middle of 2021, written English, but his actual level of competence is not known due to lack of testing.

Data were gathered through Google Takeout where YouTube data on videos watched and comments posted by the researcher's account on videos are archived and can be downloaded on request from Google. The comments were posted under the researcher's account since the participant himself has no Google account. Such comments were deemed to be his upon verification as he was the only one in the household watching the videos concerned. The first comment included is dated August 17, 2021 while the last is from May 31, 2022. There are 66 comments included in the analysis. A fellow scholar who has written about the Minasbate language and colleague of the researcher was asked to validate the data gathered.

3 Data Analysis and Discussion

3.1 Written language development level

The written language development level was measured using mean length of utterance morpheme (MLUM). The average MLUM for the participant was 8.41 as shown on Table 1. This is way above the 5.22 for his age based on Rice, et al. (2010). This result must be interpreted with caution as the MLUM was intended for oral speech and an equivalent measure for written discourse is not yet available. It may be that written utterances are longer because of the different mode. It may also be because the participant did not use punctuation at all and his constructions sometimes consisted of what are termed as run-on sentences or those which lack the full stop punctuation to divide multiple thought units.

It must be noted that the participant had minimal input from home and school when it comes to the subject matter of the utterances indicated in the table, which were mostly about the gaming videos he was watching, since no one in the household or anyone in school he has interacted with had noted common interests in those games or the topics themselves. In the household, for example, his parents did not play the video games nor watched with him the subject videos. His older brothers who were at least 10 years older had different gaming interests having grown up in an earlier period. His family members, notably, also did not speak to him in English much frequently and instead used the mother tongue. His minder probably had less input to his vocabulary as she spoke Tagalog, had less proficiency in and seldom used English, being unable to graduate from high school herself and did not distinguish herself academically. In school or neighborhood, similar-aged children did not share his interests in the games or videos watched. Thus, he had minimal input from these sources. If at all, the participant likely got his vocabulary from the other videos he has watched like those about vehicles and other games. He also does not get significant input from traditional television as the household seldom subscribed to such media. It may be said YouTube acted as the babysitter for the participant most of the time he was with his minder, a phenomenon already noted in the literature (Beyens and Eggermont, 2014).

It should be noted that the participant only started posting comments when he was already a beginning reader. He, however, was not yet proficient or even had basic skills in using a pen to write on paper. This is noteworthy as the child first wrote digitally instead of traditionally using a pen and paper.

Utterance	MLUM	Type of sentence
Part 2 pls part 2 pls or else	8	N/A
Build a bunker not that hard	6	Simple

I realized something if u use code ghost u get billion steps	13	Complex
Ice scream 6 coming soon they waited for ice scream 6	13	Simple
Dumb centi he miss one marker in the rust room	10	Simple
He accidentally put b to say gift tag	10	Simple
His friend broke every block	5	Simple
How did bedrock break	4	Simple
Crash frontier is the best	5	Simple
My name in roblox is derpypatter	6	Simple
Rules donts touch lightsabers	6	Simple
This the sodor fallout song	5	Simple
Literally thats not cool u make that jump	10	Complex
Why did he add a image of krabby patty mobile	10	Simple
Dude someone deleted Minecraft on my phone	8	Simple
Pls make video of tanks in teardown war multiplayer	10	Simple
Ay no bad word	3	N/A
I like when eystreem turned on shaders for no reason	12	Complex
When u see blood golem disconnect	7	Complex
Sonic hacked u i feel bad for jack	9	Complex
Blood golem is real it destroyed all the files	11	Complex
Bad fact i hate sebee he's bullying people	10	Complex
Yay this helped me alot I got all the markers	12	Complex
That tapwater cheats on bedwars	7	Simple

Here tips to survive when you hear a screaming run away	13	Complex
When he sees a good drawing 1 star why its a good drawing	14	Complex
Unsub pls unsub from him because of trolling	9	Complex
Unsub from hes the worst youtuber	9	Complex
I like when he sees a guy with katana he thinks hes cheating	14	Complex
I feel so bad for unspeakable that is hes has no money	12	Complex
Don't look away stupid	4	Simple
That would been weird if flashlight says stop the fucking train	13	Complex
Glitch is dumb hes accidentdldy press the black arrow	11	Complex
Press the space button to activate the balloon but u jump off	11	Complex
I think hes hacking	6	Complex
Uhh pls owner of this game delete this roblox game	7	Simple
JISHA JAGADEESH how do get rid of that oof deathsound it still plays i would never play it	17	Complex
Dislike him i dontlike him	5	Complex
OH SHIT OH SHIT	2	N/A
What is point of this	5	Simple
Dont look away from that scp peanut	7	Simple
Beep beep beep beep	4	N/A
Uhh why did he make he this videos	7	Simple
Why didn't he show us fight the ender dragon	10	Simple
Wow he not talking	4	Simple
What u guys speak Korean	6	Simple
Bacon	1	N/A
Getting on top of the plane is scary thats a dumb idea	5	Complex

Wow someone really wants to win in fortnite	9	Simple
How did he didn't know how to play fnf	9	Simple
I HATE CENTIDENT	3	Simple
Why didn't he show me beat the ender dragon	10	Simple
Centi is dumb he didn't pick up the gold helmet	11	Complex
How do u get those skins	6	Simple
How did the train get going did a rock hit the lever to get the train get going?	13	Complex
wow this guy loves teardown	5	Simple
Yeah train cannot go on sharp curves	8	Simple
U got caught for speeding Why did he race	10	Complex
This a cool song ngl	5	Simple
How does spycakes dont know how get rid of curse all do u is drink milk how ther dumb	19	Complex
Why is he watching cry hes cry all night	10	Complex
Dont ride a helicopter kids it dangerous	10	Complex
This guy is a hacker	6	Simple
I like glitch keeps banging on his table	10	Complex
Why is there a panda fish there's fish	6	Complex
And why is the ball cleaning the teeth	9	Simple
Mean	8.41	

Table 1: Written utterances and their length and sentence type

3.2 Complexity of utterances generated

The complexity of the utterances was determined by individual inspection and categorized either as simple or complex, following the criteria that having one independent clause counts as a simple sentence and having more counts as a complex sentence (Lust, Foley, and Dye, 2009). There were 29 sentences classified as complex, 32 as simple, and five which were phrases only. This means 47% of the utterances had complex syntax and higher than the 20% achieved typically by children with at least 5.00 MLU (Schuele and Dykes, 2005). It may also mean a sequential bilingual development in mother tongue and English, with the prospect of the child growing more proficient in his L2 rather than his L1.

It can be noted that the utterances were English-only, likely suggesting that for the domains of online gaming and video-watching, the participant preferred English, most probably because the content for these materials are also English-only, at least those he has watched.

Among the morpho-syntactic patterns observed include the developing use of the present, past, and progressive forms of the verb, use of the auxiliaries “do” and “did” in multi-word verbs, use of shortened constructions like “pls” and “u”, use of the imperative, declarative, and interrogative types of sentences, and non-use of the full stop, question mark punctuation for questions, and apostrophe.

3.3 Communicative purposes used

Communicative purpose was measured by inspecting and analyzing each of the utterances individually and categorizing the function used as judged by the researcher. The participant used a variety of communicative purposes – 12 different purposes of which giving verdict, asking question, and expressing disagreement were the most frequently used. Table 2 and Figure 1 together present a concerning picture of how the participant is deploying his linguistic resources. The vocabulary contains words that denote negativity and could reflect some of the disturbing behavior online and in gaming that give parents’ pause on utilizing technology and its influence in shaping children’s development. For example, parents generally view screen media as problematic and their use by young children as potentially dangerous (Mavoa, Gibbs, and Carter, 2017).

According to Austin’s theory (1975), there are three components in a speech act: “the locutionary act, or the act of saying a sentence that makes sense and refers to something; the illocutionary act, or the speaker’s purpose in saying that sentence; and the perlocutionary act, or the effect of that sentence on a listener.” For this study, inspecting the participant’s YouTube comments makes it clear the first two elements are present – his comments are clearly referred to some context in gaming and/ or videos being watched and evidently purposeful. However, the third component cannot be readily examined since there was little interaction like replies and likes/ dislikes on his posted comments. It may be because of the profile picture showing was that of the researcher's account and not his own. As he was watching the videos, he was not actively interacting with another player or interactant as he was just watching on YouTube and not playing. However, it appears from his utterances that he wanted someone to reply or he was addressing the comments to someone, likely the video maker or the others watching the videos (for example: "Pls make video of tanks in teardown war multiplayer"; "Yay this helped me a lot I got all the markers").

Communicative Purpose	n
Give verdict	14
Ask question	13
Express disagreement	11
Praise	8
Give information	7
Express emotion	6
Tell story	5
Give order	5
Make request	3
Give opinion/ make comment	2
Give warning	2
Express agreement	1

Table 2: Communicative purposes used

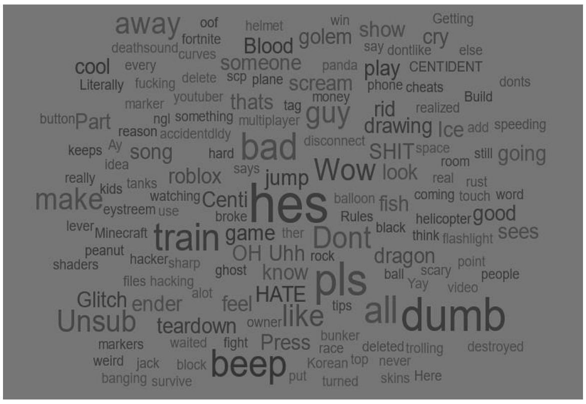


Figure 1: Word cloud generated for participant's vocabulary

3.4 Implications for Further Language and Literacy Development

It appears the participant is developing according to norm or even exceeding it if his language use in English is the basis. There may be some concern with regards to the vocabulary used which reflect those which he is exposed to. Language and literacy development other than in the mother tongue or concurrently with it is very much possible and supported in part by the results of this study.

This study may also have some implications for the language policy in place in the country. The vast sharing in language and culture due to the internet has resulted in the weakening of the distinction between a second and a foreign language (Oxford, 2017). Early exposure to second and subsequent languages could also conceivably narrow the divide between L1 and L2. As shown in the present study, the participant's language development in a second language to which he had early exposure (right around or before the age of three) may result in his attaining development comparable to first language speakers in that L2. The common definition of first language/mother tongue "as any native language developed before the age of three" (Oxford, 2017), citing Dewaele and Pernelle (2015) is similar to the legal definition under Philippine law (Congress of the Philippines, 2013) which is "language or languages first learned by a child, which he/she identifies with, is identified as a native language user of by others, which he/she knows best, or uses most". However, this presupposes languages as distinct and separate from one another and L1s being tied to ethnic identities. Languages have been argued to be "ever-developing resources" with "learners actively transforming their linguistic world" (Larsen-Freeman, 2015). While Larsen-Freeman talked about what she termed as second language development, it is not a reach to say L1s and L2s (like English in the Philippines) have a more interactive dynamic. There is acknowledgment of the Filipinos' adaption of English "to mirror and reflect their lived experiences" (Kirkpatrick, 2018). Thus, a multilingual approach reflecting these experiences of an individual may be considered.

4 Conclusion and Recommendations

The use of mean length of utterance to analyze written YouTube comments is explored in this study to determine a young bilingual's language and literacy development. Further study could be done to see whether such novel usage could be tenable. Basic measures such as syntactic complexity based on type of sentence and counting of usage of communicative purpose were used to shed light on the research questions. The rather fundamental analysis nevertheless revealed key insights which have practical implications at the familial level as well as for language policy. Parents and educators may gain some important understandings based on the results and analysis of this study.

Acknowledgments

The author thanks his research director Dr. Sherwin E. Balbuena for validating the data, the anonymous reviewers and his professor Dr. Virna Villanueva for their thoughtful and constructive comments, and the conference chair Dr. Shirley Dita, as well as the conference organizers, for this wonderful opportunity.

References

Andy Blunden. 2011. Vygotsky's Idea of Gestalt and its Origins. Theory and Psychology, 21(4): 457-471.

Andy Kirkpatrick. 2018. English in Multilingual Settings: Features, Roles and Implications. In I. PefiancoMartin (Ed.), Reconceptualizing English Education in a Multilingual Society (pp. 15-28). Springer.

Barbara C. Lust, Claire Foley, and Cristina D. Dye. 2009. The First Language Acquisition of Complex Sentences. In E. L. Bavin (Ed.), The Cambridge Handbook of Child Language (pp. 237-258). Cambridge University Press.

Colin Lankshear, Ilana Snyder, and Bill Green. 2000. Teachers and Technoliteracy: Managing Literacy, Technology and Learning in Schools. Sydney: Allen and Unwin.

Colin Lankshear, and Michele Knobel. 2011. New Literacies: Everyday Practices and Classroom Learning (3rd ed.). Maidenhead: Open University Press.

Congress of the Philippines. 2013. Enhanced Basic Education Act of 2013. Retrieved from https://www.lawphil.net/statutes/repacts/ra2013/ra_10533_2013.html

David Barton, and Carmen Lee. 2013. Language Online: Investigating Digital Texts and Practices. London: Routledge.

Diane Larsen-Freeman. 2015. Saying What We Mean: Making a Case for "Language Acquisition" to Become "Language Development". Language Teaching, 48(4), 491–505.

Gunther Kress, and Theo van Leeuwen. 2001. Multimodal Discourse: The Modes and Media of Contemporary Communication. London: Arnold Publishers.

Hansun Zhang Waring. 2018. Discourse Analysis: The Questions Discourse Analysts Ask and How They Answer Them. Routledge.

Helen Kelly-Holmes, and Tommaso Milani. 2013. Thematising Multilingualism in the Media. Philadelphia: John Benjamins.

Ilana Snyder and Ekaterina Tour. 2017. Research Approaches to the Study of Literacy, Learning, and Technology. In Kendall A. King, Stephen May, and Yi-Ju Lai (Eds.), Research Methods in Language and Education (3rd ed.). Springer, Cham, Switzerland.

Ine Beyens and Steven Eggermont. 2014. Putting Young Children in Front of the Television: Antecedents and Outcomes of Parents' Use of Television as a Babysitter. Communication Quarterly, 62(1), pp. 57–74.

Ingrid de Saint-Georges. 2017. Researching Media, Multilingualism, and Education. In Kendall A. King, Stephen May, and Yi-Ju Lai (Eds.), Research Methods in Language and Education (3rd ed.). Springer, Cham, Switzerland.

Jack R. Fraenkel, Norman E. Wallen, N., and Helen Hyun. 2011. How to Design and Evaluate Research in Education. McGraw-Hill.

Jackie Marsh. 2011. Young Children's Literacy Practices in a Virtual World: Establishing an Online Interaction Order. Reading Research Quarterly, 46(2), 101–118.

James Paul Gee. 2003. What Video Games Have to Teach Us about Learning and Literacy. New York: Palgrave Macmillan.

Jane Mavoa, Martin Gibbs, and Marcus Carter. 2017. Constructing the Young Child Media User in Australia: A Discourse Analysis of Facebook Comments. Journal of Children and Media, 11(3), 330-346. https://doi.org/10.1080/17482798.2017.1308400

Jean-Marc Dewaele, and Lorette Pernelle. 2015. Emotion Recognition Ability in English Among L1 and LX Users of English. International Journal of Language and Culture, 2(1), 62–86. https://doi.org/10.1075/ijolc.2.1.03lor

John L. Austin. 1975. How to Do Things with Words. Harvard University Press.

Judith B. Bryant. 2009. Pragmatic Development. In E. L. Bavin (Ed.), The Cambridge handbook of child language (pp. 339-354). Cambridge University Press.

Kenneth Reeder, Jon Shapiro, Rita Watson, and Hillel Goelman (Eds.). (1996). Literate Apprenticeships: The Emergence of Language and Literacy in the Preschool Years. Ablex.

Lev S. Vygotsky. 1978. Mind in Society: The Development of Higher Mental Processes. Harvard University Press.

Mark Warschauer. 2006. Laptops and Literacy: Learning in the Wireless Classroom. New York: Teachers College Press.

Max Coltheart. 2005. Modeling Reading: The Dual-Route Approach. In M. Snowling, and C. Hulme (Eds.), The Science of Reading: A Handbook (pp. 6-23). Blackwell.

Melanie Schuele, and Julianna C. Dykes. 2005. Complex Syntax Acquisition: A Longitudinal Case Study of a Child with Specific Language Impairment. Clinical Linguistics and Phonetics, 19, 295–318. doi: http://doi.org/10.1080/02699200410001703709

Nicholas C. Burbules, and Thomas. A. Callister, Jr. 2000. Watch IT: The Risks and Promises of Information Technologies for Education. Boulder: Westview Press.

Pamela Takayoshi, Emily Huot, and Meghan Huot. 1999. No Boys Allowed: The World Wide Web as a Clubhouse for Girls. Computers and Composition, 16(1), 89–106.

Patrice L. Engle, Maureen M. Black, Jere R. Behrman, Meena C. de Mello, Paul J. Gertler, Lydia Kapiriri, Reynaldo Martorell, Mary E. Young, and International Child Development Steering Group. 2007. Strategies to Avoid the Loss of Developmental Potential in More Than 200 Million Children in the Developing World. The Lancet, 369(9557), 229-242. Retrieved from https://digitalcommons.calpoly.edu/cgi/viewcontent.c gi?article=1004&context=psycd_fac

Rebecca L. Oxford. 2017. Conditions for Second Language (L2) Learning. In N. V. Deusen-Scholl, and S. May (Eds.), Encyclopedia of Language and Education: Second and Foreign Language Education (3rd ed., pp. 27-42). Springer.

Rémi A. van Compernolle. 2022. A Sociocultural Theory Perspective on Sociolinguistic and Pragmatic Variationin L2 Development. In Kimberly Geeslin (Ed.), The Routledge Handbook of Second Language Acquisition and Sociolinguistics. Routledge, New York, NY and London, UK.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., and Blossom, M. 2010. Mean Length of Utterance Levels in 6-Month Intervals for Children 3 to 9 Years with and Without Language Impairments. Journal of Speech Language and Hearing Research, 53(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183

Roger Brown. 1973. A First Language. Allen and Unwin.

Sherry Turkle. 1995. Life on the Screen: Identity in the Age of the Internet. Simon and Schuster, New York, NY.

Sonia Livingstone. 2002. Young People and New Media: Childhood and the Changing Media Environment. London: Sage.

Sonia Livingstone. 2012. Critical Reflections on the Benefits of ICT in Education. Oxford Review of Education, 38(1), 9–24.

Metalinguistic Corrective Feedback and Students’ Response to Feedback in L2 Writing

Pilar S. Caparas, Ph. D.
Western Mindanao State University
Region IX, Zamboanga City
pilarcaparas@gmail.com

Abstract

The study of corrective feedback (CF) has been gaining more prolific attention in the field of Second language Acquisition (SLA) up to this time. Theorists, researchers, educators have been investigating which forms of CF are effective. The study determined the use of metalinguistic corrective feedback and students’ response to such feedback in L2 writing. It also investigated the students’ belief on which features of their writing teachers should pay attention to. Forty students (8=males, 32=females) were given two writing tasks in which the teacher coded the errors using metalinguistic clues in their essays. The students revised their first draft following the given linguistic error codes. From these tasks, the teacher analyzed the preponderant linguistic errors and the types of revisions the students incorporated employing the list of error codes adopted from Corpuz (2010), Ellis’s (2009) typology of written corrective feedback and Ferris’s (2006) types of revisions. The results showed that the most prevalent linguistic errors committed by the students were *punctuation* and the most common type of revision category in the redrafts of the students was *error corrected*. Further, the survey conducted to the students revealed that *error in grammar (82%)* is the top most feature the students would like their teacher to correct in their writing. This seemed to suggest that the students have positive perception on corrective feedback. The results suggested that during the revision, the metalinguistic clues were not enough to rectify the errors and that the linguistic competence of the students was needed to make the corrections. Essentially, most of the participants strongly agreed that the use of metalinguistic corrective feedback with the use of correction symbols facilitate their revision tasks.

Keywords: *Corrective feedback; Metalinguistic corrective feedback, grammatical errors, SLA*

1 Introduction

Corrective Feedback (CF) has been a prolific ground for research up to this time. “Corrective feedback refers to any signal that a learner’s utterance may be erroneous in some way” (Nassaji & Kartchava, 2021, p.1). “It is also a response that is provided by a teacher, a researcher, or a peer in reaction to an error committed by the second/foreign language (L2) learner”(Leow & Driver, 2021, p. 65) A CF can be oral or written.

The role of written corrective feedback (WCF) has been a topic of immense interest in Second Language Acquisition (SLA) research to date, according to Brown (2007). Lightbown and Spada (2009) broadly define CF, also known as negative feedback, as “any indication to the learners that their use of the target language is incorrect.” Van Beuningen (2010) affirms that “CF is a widely applied pedagogical tool and its use finds support in SLA theory, yet practical and theoretical objections to its usefulness have been raised” (p.1). To extend this definition to the written discourse, written corrective feedback (WCF), refers to “various ways a reader can respond to a second language writer by indicating that some usage in the writing does not conform to the norms of the target language” (Sun, 2013, p.1). The literature on corrective feedback also received several reviews particularly on their roles in the L2 class, and researchers were interested “if and how CF can help students to become able and self-employed writers” (Van Beuningen, 2010, p2.)

The present study analyzed the metalinguistic CF coded in the L2 writing of high school students. It determined how the learners responded to the corrections provided. The learner’s response frequently took the form of revision of the initial draft – an important stage in process writing. Much of the research that has investigated written CF has also centered on whether learners are able to make use of the feedback they receive when they revise. Thus, the study also investigated this aspect. It looked into the types of revisions the students used in their drafts which the teacher marked using metalinguistic clues and identified features of the students' writing that they believed teachers should pay attention to.

1.1 Review of Related Literature

Studies on written CF were conducted in various settings such as in the classroom, in computer mediated communication, in training and even in areas dealing with students with disabilities. Its types of feedback were also found to produce different effects and results. One of the main findings of research on written corrective feedback was that CF was helpful in facilitating L2 writing of students but the lack of knowledge on the rules of grammar make CF counterproductive (Ferris, Liu, Sinha, and Senna, 2013; Sauro, 2009). Essentially, in the study conducted by Van Beuningen (2010), she concluded that by offering learners opportunities to notice the gaps in their developing L2 systems, engaging in metalinguistic reflection, written CF has the ability to foster SLA and lead to accuracy development.

Furthermore, it was found that the use of written CF is strengthened when it is followed by a teacher conference and peer-peer interaction than receiving only CF and teacher conference (Chuang, 2009). On the one hand, there were also factors which influenced CF as found in the study of Ferris, Liu, Sinha & Senna (2013). The analysis in this study was focused primarily on the students' description of their own self-monitoring processes as participants revised marked papers and wrote new texts. Individuals and contextual factors appeared to influence their writing development.

On the types of feedback, Shirazi and Shekarab (2014) investigated the effect of direct and indirect feedback on Iranian learners studying Japanese language. They found that the group which received direct and indirect CF every other session had higher mean than the group that received only direct feedback. Further results showed that direct CF had little or no role to play in the writing practices of the group that received it. This finding seemed to strengthen the result of the study conducted by Parreno (2014) which suggested that using coded corrective feedback was a better approach than direct correction or indirect correction, although its efficacy on second language learning/acquisition needed further investigation.

Moreover, McNulty (2007) found out that recasts was the most commonly used feedback type, yet it was least effective in terms of student uptake, while the most successful feedback are repetition, metalinguistic, elicitation and clarification which were least used by the teachers.

In addition, more than investigating which among the types of CF is more effective, the belief of the learners on CF must also be investigated for CF to be useful. Lennane (2007) in his descriptive analysis on the preferences for types of errors to correct and effective reactions to error correction found that explicit correction ranked the highest followed by recasts and then prompts. Diab's (2006) study shed light on an important role in providing feedback. She recommended that teachers should incorporate classroom discussions on error to help their students understand how feedback is intended to affect their writing and why it is given in a particular way.

These studies supported the present study. Indeed, there are much more significant features of written CF that still can be explored through scholarly research.

1.3 Research Questions

The present study sought to answer the following questions:

1. What are the most preponderant linguistic errors committed by the high school students in L2 writing?
2. What types of revisions do the students incorporate in their texts when they were provided with corrected feedback in the form of metalinguistic clues?
3. Does receiving written corrective feedback facilitate the linguistic accuracy of L2 writing among the students?
4. What features of students' writing do they believe are the most important for their teachers to correct?

1.4 Theoretical Framework

The potential benefits of employing written corrective feedback to language learning depend on various theoretical grounds. The present study is anchored on the following theoretical frameworks/models.

1.4.1 On Corrective Feedback

Ellis (2009) identifies a typology of teacher options or strategies for correcting students' written work. He focuses on one kind of correction which is the correction of linguistic errors**.**

1.4.1.1 Direct CF

In this strategy, the teacher provides the students with the correct form.

1.4.1.2 Indirect CF

In utilizing this strategy the teacher indicates that an error exists but does not provide the correction. This type of CF can be (a) indicating + locating the error and (b) indicating only. The former takes the form of underlining and use of cursors to show omissions in the student's text while the latter takes the form of an indication in the margin that an error or errors found in a line of the text.

1.4.1.3 Metalinguistic CF

In this strategy, the teacher provides some kind of metalinguistic clues as to the nature of the error. The teacher can use two types of metalinguistic clues: (a) Use of error code and (b) Brief grammatical descriptions. The former has the teacher write codes in the margin or above the location of the error while in the latter has the teacher number the errors in text and writes a grammatical description for each numbered error at the bottom of the text.

1.4.1.4 The Focus of the Feedback

This concerns whether the teacher attempts to correct all (or most) of the student's errors or selects one or two specific types of errors to correct. This distinction can be applied to each of the above options. This type can be (a) unfocused CF and (b) focused CF.

1.4.1.5 Electronic Feedback

The teacher indicates an error and provides a hyperlink to a concordance file that provides examples of correct usage.

1.4.1.6 Reformulation

This consists of a native speaker's reworking of the students' entire text to make the language seem as native-like as possible while keeping the content of the original intact.

The present study employed the metalinguistic CF using error codes.

1.4.2 On Student's Response to Feedback

Ellis (2009) provides the typology of student's response to feedback as follows:

1.4.2.1 Revision Required

1.4.2.2 No Revision Required

This can take the forms of (a) students asked to study corrections and (b) Students just received corrected text. In this study, the teacher required the students to do a revision following the error codes marked on their texts.

1.4.3 CF as a Focus-on-Form Intervention

Focus-on Form approach by Long (Long 1991; 1996; 200; Long & Robinson, 1990, as cited in Van Beuningen, 2010), is a pedagogical intervention that has received considerable attention and which has been advocated in the SLA.

According to Long (2000, p. 85 in Van Beuningen,2010, p. 4), focus on form "involves briefly drawing students' attention to linguistic elements […] in context as they arise incidentally in lessons whose overriding focus is on meaning or communication. The temporary shifts in focal attention are triggered by students' problems with comprehension or production." One of the most crucial characteristics of a focus-on-form intervention is that it is provided within a communicative context. Long implied that focus– on form episodes are unplanned (i.e. incidental). This implication had contrasted the definition of other scholars.

1.5 Conceptual Framework

Following the theoretical underpinnings of the present study, the researcher focused on the use of the metalinguistic CF. Metalinguistic CF involves providing learners with some form of explicit comment about the nature of the errors committed. Lyster and Ranta (1997 as cited in Rezaei, 2011) categorize metalinguistic feedback as "comments, information, or question related to the well-formedness of the student's utterance, without explicitly providing the correct form" (p. 657). Metalinguistic comments, the most minimally informative method than recasts, simply indicate the occurrence of an error. The metalinguistic CF which the present study is referring to here is the metalinguistic unfocused CF in which the researcher labels the errors of the students using error codes. An example from the data is shown below.

P C
I agree. Because in our family my parents are
S/V
the one who supports us.

Following the error codes, the student will do the revision as follows:

I agree because in our family my parents are the ones who support us.

Furthermore, CF can be focused or unfocused corrective methodologies. The present study employed the unfocused approach which involves correction of all errors in a learner's text, irrespective of their error category.

For feedback to work for either redrafting or language learning, learners need to attend to the corrections (Ellis,2009, p.99.). The taxonomy by Ferris (2013 as cited in Ellis,2009, p. 105) was used by the researcher to determine the learners' response to the feedback.

The study employed a descriptive research design that utilized writing tasks in the form of essay writing and the written output of the students received metalinguistic clues. The representation of the overall conceptual research design is shown below.

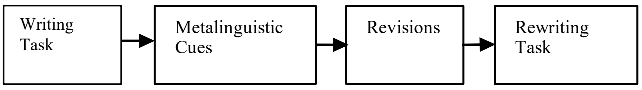


Figure 1. Conceptual research design

2 Methodology

2.1 Research Design

Given the nature of the investigation, the present study used the descriptive qualitative approach. Frequency counts and percentages were also used to determine the number of errors that occurred during the writing, to determine the frequency of the types of revisions the participants incorporated in their responses to the coded error; and to determine the preponderant features of the students' writing the students believed the most important for their teacher to correct.

2.2 Setting

The setting of the study is an accredited private high school in Bulacan. The administrators and faculty are really working to produce quality instruction to respond to the needs of times. The school is a prominent school in the province where most of the students are considered well-off. The teachers are mostly new, but they are under the close supervision of the academic coordinator and school directress. The writing sessions were done through the English subject.

2.3 Participants

Forty (male=8; female=32) private high school students enrolled for the school year 2014-2015 participated in the present study. They belonged to the first section of the graduating class. Their ages ranged from 15 to 17 years. Their final accumulated grades in the third year ranged from 80 to 96.

2.4 Data Collection

Data collected by the researcher were primarily through the written output of the students. The researcher administered two writing sessions and two sessions for the students to revise their corrected essays. For additional data, the researcher administers a brief survey to determine which features of the students' writing they believe to be important for their teacher to correct.

2.5 Data Collection Procedure

Approval from the school head was sought primarily to conduct data gathering. The approval was given, and the school head asked the researcher to coordinate with the English teacher of the fourth year students. The researcher was facilitated by the English Language teacher to conduct the writing and revision sessions which took place in a week.

The participants were given two writing tasks in the form of essay writing. The students accomplished each writing task in thirty minutes.

After each writing task, the researcher checked the writing by indicating the error codes above the location of the error in the texts of the students. The teacher gave back the coded written output on the same day for the students to do the revision. The participants also did the revision or rewriting for thirty minutes. A list of error codes was provided to the participants during each revision. The Error codes, which the researcher used, were adopted from the study of Corpuz (2010). The revised written output and the original written output were collected again.

In the coding stage, the researcher listed all the errors and classified them according to linguistic errors based on the list of error codes.

The researcher counted the number and frequencies of these errors to determine the most preponderant occurrences of each. The linguistic errors are: use of wrong word, missing word, punctuation, capitalization, tense, word form, subject-verb agreement, plural/singular, spelling mistake, preposition, word order, article use, extra word, cannot be understood sentences, register, active/passive, awkward sentence and pronoun use.

Moreover, the researcher also counted the types of revisions the participants incorporated in their drafts. These types of revisions were based on the taxonomy of Ferris (2013. in Ellis, 2009, p. 105), which are : error corrected (Error corrected per teacher's marking), incorrect change (Change was made but incorrect), no change (No response to the correction was apparent), deleted text (Participants deleted marked text rather than attempting correction), substitution, correct (Participants invented a correction that was not suggested by the teacher's marking) and substitution, incorrect (participants incorrectly made a change that was not suggested by the teacher's marking). The two essay topics given by the researchers were (a) Describe Your Hometown and (b) Do you agree with the statement that parents are the best teachers.

The researcher also conducted a brief survey to get additional data on the perceived belief of the participants as to which of the features of their writing they believed were the most salient for the teachers to correct to. Their responses were also subjected to frequency counting.

2.6 Method of Data Analysis

To answer research questions 1 to 4, the researcher used frequencies and percentages to determine the preponderant occurrences. These frequencies were the result of the coding done primarily with the data using the typologies used in the study.

3 Results and Discussion

3.1 Introduction

The study delved on the metalinguistic CF marked by the teacher on the essays of the high school students and what types of revisions the students employed on these coded errors on their essays.

3.1.1 On the Types of Linguistic Errors Committed by the Participants

Table 1 shows the types of linguistic errors the students committed in writing their essays.

Table 1
Linguistic errors committed by the students

	Linguistic Errors	First Essay		Second Essay		Total Errors	
		f	%	f	%	f	%
1	Punctuation	241	21.71	97	17.60	338	20.34
2	Wrong Word	155	13.96	54	9.80	209	12.58
3	Awkward	126	11.35	43	7.80	169	10.17
4	plural/ singular	83	7.47	81	14.70	164	9.87
5	Cannot be understood	83	7.47	29	5.26	112	6.74
6	Missing word	62	5.58	48	8.71	110	6.62
7	Capitalization	44	3.92	56	10.16	100	6.02
8.	Word form	78	7.02	20	3.62	98	5.90
9.	Subject-verb agreement	30	2.70	42	7.62	72	4.33
10	Preposition	43	3.87	23	4.17	66	3.97
11	Tense	41	3.69	23	4.17	64	0.03
12	Pronoun use	37	3.33	18	3.26	55	3.31
13	Article use	41	3.69	11	1.99	52	3.13
14	Spelling mistake	18	1.62	4	0.72	22	1.32
15	Word order	16	1.44	2	0.36	18	1.08
16	Register	9	0.81	0	0	9	0.54
17	Active/passive	3	0.27	0	0	3	0.18
	Total	1,110	100	551	100	1,661	100

The data recognized 1,661 occurrences of different linguistic errors. Table 1 shows that the most prevalent committed error by the students is punctuation in which the occurrences is one-fifth of the total occurrences recognized in the data. The results also reveal that there were slim proportions of other linguistic errors committed by the participants during the writing of essays such as wrong word- 209 (12.58%) occurrences, awkward sentence – 169 (10.17%) occurrences and plural/singular -164 (9.87%) occurrences. The very least linguistic errors are register- 9 (0.54%) and active/passive -3 (0.18%) occurrences. It is worthwhile to note that the participants seemingly do not have sufficient mastery of the use of comma, semicolon, apostrophe and period. The results also suggest that the participants have difficulty in vocabulary as revealed in their use of wrong words and wrong register and lack of clarity in written expressions as they produced awkward sentences. They do not also show accuracy in the use of plural and singular forms of the words.

Furthermore, it can be observed also that the participants' linguistic errors in the second essay had decreased tremendously. This can be accounted for by the researcher's observation that the students became conscious of their writing during the administration of the second essay.

3.1.2 On the Types of Revisions Performed by the Students

The data recognized a total of 1,574 occurrences of different revisions performed by the participants during the revision. This number is smaller than the number of linguistic errors marked by the teacher. This can be accounted for by the way a participant revised his or her coded essay. A long stretch of sentence having five linguistic errors, for instance, can be deleted by a participant during the revision. Thus, reducing the number of the types of revisions used by the participants.

When the participants rewrote their coded essays, fifty percent of the total revisions show that they corrected the errors appropriately while an accumulated percentage of less than fifty percent of the revisions performed are incorrect, deleted, incorrectly substituted and ignored or unchanged and only seven percent shows correct substitution for awkward sentences as shown in Table 2.

Table 2

Types of revision incorporated by the participants

Types of Revisions	First Essay		Second Essay		Total	
	f	%	**f**	%	**f**	%
Error Corrected (EC)	522	46.52	229	50.66	751	47.71
Incorrect Change (IC)	176	15.68	71	15.70	247	15.69
Deleted Text (DT)	167	16.01	44	9.73	211	13.40
Substitution Incorrect (SI)	136	14.88	35	7.74	171	10.86
No Change (NC)	76	6.77	41	9.07	117	7.43
Substitution Correct (SC)	45	4.01	32	7.07	77	4.89
Total	1,122	100	452	100	1,574	100

3.1.3 Does receiving written corrective feedback facilitate the linguistic accuracy of L2 writing among the students?

Figures 2 and 3 show the errors corrected and incorrect changes made by the students.

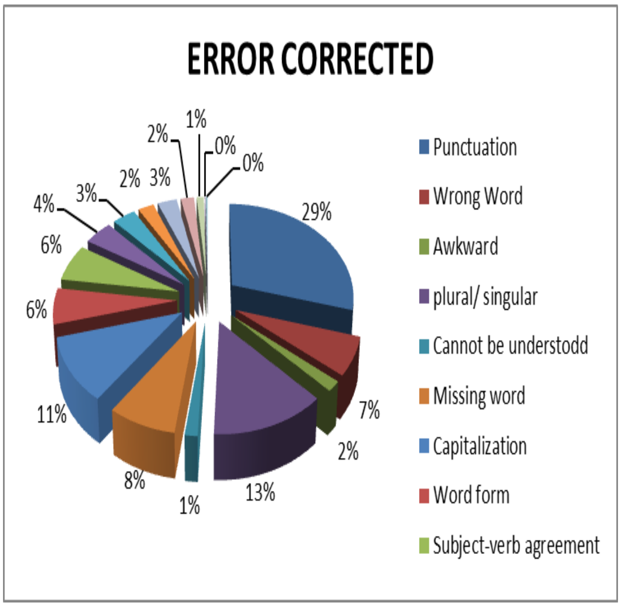


Figure 2. Percentages of Errors Corrected

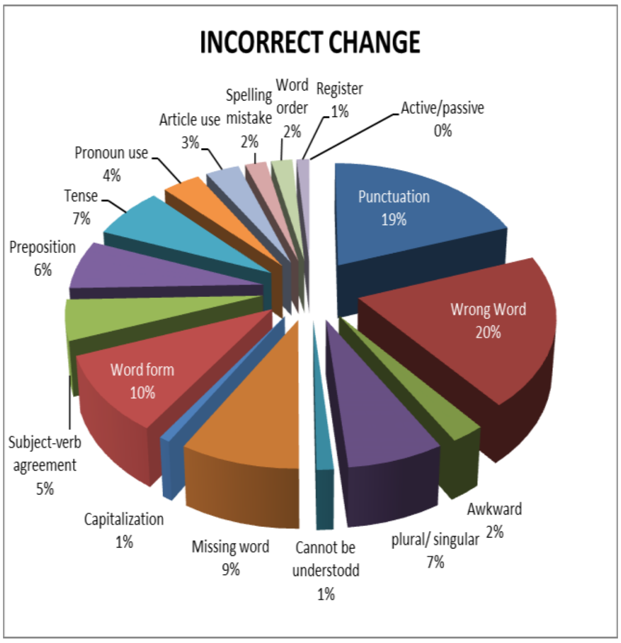


Figure 3. Percentages of Incorrect Changes

In order to answer this research question, the researcher looked closer to the types of revision done by the participants on each of the linguistic errors.

The results show that error corrected type of revision is the most preponderant type of revision. It occurs less than fifty percent of the total types of revisions recognized in the data. Its preponderant use is evident in greater occurrences in punctuation, wrong word, plural/singular, missing word, capitalization, word form, subject-verb agreement, preposition, tense, pronoun use, article use, spelling mistake, word order and the use of active and passive than in awkward, cannot be understood sentences and register. This further implies that the participants have difficulty correcting linguistic errors appropriately on awkward and cannot be understood sentences and register as shown in Figure 2.

The results also show that the second most preponderant type of revision incorporated by the participants is incorrect change that have instantiations of 247 (15.69%). The linguistic errors which have the most incorrect changes are punctuation, word order, plural/ singular, missing word, word form, preposition and tense while the linguistic errors which have the least incorrect changes are awkward and cannot be understood sentence, capitalization, subject-verb agreement, pronoun, article, spelling mistake, word order and register as shown in Figure 3. Incorporating incorrect revision with a slim percentage implies that the student found a slight difficulty in correcting their errors during the revision.

Furthermore, the third most preponderant types of revision done by the participants is deleted text followed by incorrect substitution, no change and very slim percentage of substitution correct. In summary, Table 3 shows errors corrected by the participants and substitution corrected only mark 52.60% while incorrect change, no change, deleted text and substitution incorrect mark equivalent to 47.38%. These results suggest that the students can correct errors but at the same time also lack the linguistic competence to rectify errors. The errors that they found most difficult to correct during revision are correcting awkward sentences and sentences that cannot be understood and registered. During the revision, some of these were deleted and incorrectly substituted.

Moreover, the question as to whether metalinguistic CF facilitates the writing of the students is explained by the previous discussions of the result. In other words, receiving written corrective feedback in the form of metalinguistic clues may lead participants to see the nature of their error and help them produce the corrections, but this does not warrant that their corrections are appropriate. The participants also need to apply their linguistic competence to rewrite difficult errors (e.g. awkward, cannot be understood sentences and register) appropriately. If they did not develop, in all likelihood, they will commit errors and at the same time cannot rectify the errors. The results are in consonance with the findings of Sauro (2009) when she investigated the impact of two types of computer mediated corrective feedback in the form of recasts and metalinguistic information on the development of adult learners' L2 knowledge. Sauro (2009) found no significant advantage for either feedback type on immediate sustained gains in knowledge of target forms, although the metalinguistic group showed significant immediate gains relative to the control condition. The longitudinal study conducted by Ferris, Liu, Sinha and Senna (2013) found that students found the techniques used in the study (focused WCF, revision, and one-to one discussion about errors) useful, but formal knowledge of language rules played a limited and sometimes even counterproductive role in their self editing and composing.

3.1.4 Features of the students' writing they believe are the most important for their teachers to correct

Table 3

Writing features students believe teachers Should correct

Writing Features	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree	Total
a. error in grammar	28(82.35%)	5(32.78%)	1(2.94%)			34
b. error in spelling	25(73.52%)	9(26.47%)				34
c. vocabulary choice	25(73.5 2)	9(26.47%)				34
d. use of correction symbols	19(55.88)	7(20.58%)	4(11.76%)	4(11.76%)		34
e. organization of the paper	17(50%)	15(44.11%)	2(5.88%)			34
f. error in punctua-tion	15(44.11%)	13(38.23)	5(32.78%)			34
g. comments on the ideas expressed	10(29.41%)	13(38.23%)	9(26.47%)	2(5.88%)		34

Thirty-four (34) from the original 40 participants answered the survey. The results show that the participants strongly agree that teachers should point out errors in grammar, spelling, vocabulary, punctuation, and use of corrections symbols. They agree that teachers should make comments on the ideas expressed in the paper as shown in Table 3. The results imply that students have a positive attitude on the written corrective feedback provided by the teacher.

4 Conclusion

Providing written corrective feedback in the form of metalinguistic clues is a productive strategy in facilitating the L2 writing of the students. It can also be used as an assessment tool for both the teacher and the students. For the former, he or she can further design his or her instruction on the immediate needs of the students, and for the latter, they can revisit their past grammar lessons and may have a self-study approach to strengthen their linguistic competence. Using metalinguistic clues has also been beneficial for the students to be aware of the writing skills that they have, and this realization should lead them to possess some measures that can improve such skills.

4.1 Recommendation

As a teacher and researcher, I share the same recommendation in the study of Ferris, Liu, Sinha & Senna (2013, p. 307). Their findings suggested that teachers should have a more finely tuned approach to corrective feedback and that future research designs investigating written corrective feedback should go beyond consideration of students' written products only. The present study also recommends that teachers incorporate classroom discussion on error correction, feedback, and writing in order to help their students understand how feedback is intended to affect their writing and why it is given in a particular way. It is also important that teachers should become aware of their own beliefs about error correction and feedback to student writing.

4.2 Implications

3.2.1 On Pedagogy and Instruction

The results of the present study are beneficial for the institution concerned. From the results, the English area can design an effective program targeting the needs of its clientele. The school curricular revision can identify specific targets to address the needs of the students in writing, since writing is one of the most important skills a student should develop.

3.2.2 On Research

Conducting research is always part of any scholarly endeavor in his field of specialization. It is the researcher's contribution to the discipline. The results of the present study hopes to contribute ideas on the field of research particularly in the Philippine school setting. Other means of improving student writing skills should be explored by teachers in the field.

References

Chuang, W. (2009). The effects of four different types of corrective feedback on EFL students' Writing in Taiwan. Retrieved from http://cd,dyu.edu.tw/html/publication/files/JoGE04/07.pdf

Corpuz, V.A. (2010). Error correction in second language writing. Retrieved from http://eprints.qut.edu.au/49160/Victor_Corpuz_

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. Retrieved from http://sls.hawaii.edu/gblog/wpcontent/uploads/2011/08/1990-crookes utterance1.pdf

Diab, R.L. (2006). Error correction and feedback in the writing classroom comparing instructor and student preferences. *English Teaching Forum 44*(3), 1-13. Retrieved from http://americanenglish.state.gov/resources/ English-teaching-forum-2006-volume-44

Ellis, R. (2009). A typology of written corrective feedback types. Retrieved from http://lrc.cornell.edu/events/09docs/ellis.pdf

Ferris, D., Liu, H., Sinha, A. & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing 22*(1), 307-329.

Lennane, M. (2007). Cross-cultural influences on corrective feedback preferences in English Language instruction. Retrieved from http://0-search.proquest.com.lib1000.dlsu.edu.ph/ Pqdtft/docview/304344195/fulltextPDF/9F9235829A5840CAPQ/33?accountid=28547

McNulty, A. (2007). A study of corrective feedback and uptake patterns as observed in four foreign language classrooms at selected mid-western public schools. Retrieved from http://O-search.proquest.com.lib1000.dlsu.edu.ph/pqdtft /docview/304827598/ fulltextPDF/9F9235B29Ar840CAPQ/18?accountid=28647

Parreno, A.A. (2014). Student response to written corrective feedback. Retrieved from http://164.115.22.25/ojs222/index.php/LEARN/article/view/252/242

Rezaei, S. (2011). Investigating recasts and metalinguistic feedback in task-based grammar instruction. *Journal of Language Teaching and Research 2*(3), 655-663.

Sauro, S. (009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology 13*(1), 96120.

Shirazi, M. A. & Shekarabi, Z. (2014). The role of written corrective feedback in enhancing the linguistic accuracy of Iranian Japanese learner's writing. *Iranian Journal of Language Teaching Research 2*(1), 99-118.

Sun, S. (2013). Written corrective feedback: Effects of focused and unfocused grammar correction on the case acquisition in L2 German. (Order No. 3591742, University of Kansas). ProQuest Dissertations and Theses, 234.

Van Beuningen, C. (2010). Corrective feedback in L writing: Theoretical perspective, empirical insights, and future directions. *International Journal of English Studies 10*(2), 1-27.

Automatic Minuting: A Pipeline Method for Generating Minutes from Multi-Party Meeting Transcripts

Kartik Shinde[†], **Tirthankar Ghosal**[‡], **Muskaan Singh**[*], **and Ondřej Bojar**[‡]
[†]Indian Institute of Technology Patna, Bihta, Bihar, India
[‡]Charles University, Faculty of Mathematics and Physics, ÚFAL, Czech Republic
[*]Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
kartik_1901ce16@iitp.ac.in, msingh@idiap.ch
(ghosal,bojar)@ufal.mff.cuni.cz

Abstract

Automatically generating meeting minutes is a challenging yet time-relevant problem in speech and natural language processing. Nowadays, meeting minutes seem more crucial than ever due to the manifold rise of online meetings. However, *automatic minuting* is not straightforward for various reasons: obtaining transcriptions of sufficient quality, summarizing long dialogue discourse, retaining topical relevance and coverage, handling redundancies and small talk, etc. This paper presents our investigations on a pipelined approach to automatically generate meeting minutes using a BART model (Bidirectional and Auto-Regressive Transformers) trained on multi-party dialogue summarization datasets. We achieve comparable results with our simple yet intuitive method with respect to previous large and computationally heavy state-of-the-art models. We make our code available at https://github.com/ELITR/minuting-pipeline.

1 Introduction

Ever since most of our interactions went virtual, the need for automatic support to run online meetings became essential. Due to frequent meetings and the resulting context switching, people are experiencing an information overload (Fauville et al., 2021) of epic proportions. Hence a tool to automatically summarize a meeting transcript would be a valuable addition to the virtual workplace. *Automatic Minuting* is the task of generating bullet-point meeting minutes from multi-party meeting transcripts. The AutoMin shared task at Interspeech 2021 (Ghosal et al., 2021) is a community-wide effort in this direction. Organizers of AutoMin (Ghosal et al., 2022a) released a medium-scale annotated corpus (Nedoluzhko et al., 2022) of transcript-minute pairs for conducting the shared task.

Automatic Minuting is close to summarization but not the same; subtle differences exist. Summarization aims at generating a concise and coherent text summary. It often purposely removes some less critical information; minuting is more inclined towards adequately capturing the entire contents of the meeting (*coverage is probably more significant than coherence and conciseness*).

Summarizing spoken multi-party dialogues comes with challenges: incorrect or noisy automated speech recognition (ASR) outputs, long discourse, topical shifts, the dialogue turns, redundancies and small talk, etc. Hence we deem automatic minuting to be more difficult than text summarization.

Due to the variety of sub-problems associated with this task, we adopt a pipelined approach. Our method encompasses (i) pre-processing the ASR-generated meeting transcripts to drop redundancies and noise, followed by (ii) unsupervised topical segmentation, and finally (iii) summarizing each segment of the discourse with a BART model (Raffel et al., 2019) pre-trained on a large-scale dialogue summarization dataset. Our initial investigation yields encouraging results. The obtained minutes resemble the human gold standard in terms of readability and coverage. Our main contribution lies in developing a lightweight, easy-to-implement, and efficient automatic minuting pipeline by leveraging pre-trained Transformer-based language models fine-tuned on large-scale dialogue summarization datasets.

2 Related Work

Although meeting summarization is a well-studied problem in the summarization literature (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; See et al., 2017; Celikyilmaz et al., 2018; Chen and Bansal, 2018; Zhong et al., 2019; Xu and Durrett, 2019; Liu and Lapata, 2019; Lebanoff et al., 2019; Cho et al., 2019; Wang et al., 2020; Xu et al., 2019; Jia et al., 2020), automatic minuting is defined as a task relatively recently (Ghosal et al., 2021). We survey some of the relevant meeting summarization research in this section.

Early studies like Chen and Metze (2012) used intra-speaker topic modeling to summarize meetings. Later, several approaches (Zhao et al., 2019; Liu and Chen, 2019; Liu et al., 2019) documented the efficacy of hierarchical methods in learning the inherent structure of conversations. Li et al. (2019) utilized a multi-modal hierarchical attention mechanism across the topic, utterance, and word levels for the task. However, their method depends on manual annotation of topical segments and visual attention of the participants in the meetings, which are not commonly available. Zhu et al. (2020) introduced a hierarchical network *HMNet* for end-to-end training with cross-domain flexibility, which is now one of the state-of-the-art models for meeting summarization but is very resource-intensive. Recently, Liu and Chen (2021) proposed a dynamic sliding window strategy for abstractive summarization, achieving a close to state-of-the-art performance. Along similar lines, Zhong et al. (2021) presented a pre-training approach for long dialogue understanding and summarization with window-based denoising. Zhang et al. (2021) introduced a flexible multi-stage framework for longer input texts, combining a multi-stage greedy transcript segmentation with end-to-end training. Singh et al. (2021) tested several baseline text summarization models for automatic minuting and concluded that *off-the-shelf* summarization models are not suited for the concerned task.

Most of the above deep neural models are resource-heavy. The hierarchical model, HMNet, requires 4 Tesla V-100 GPUs with 32G memory on each. Our proposed pipeline approach is straightforward and consists of separate stages for each sub-task in the pipeline: pre-processing, redundancy elimination, transcript segmentation, summarization, and post-processing. Each stage has a unique problem, with specified target outputs, culminating in the final objective, i.e., minutes generation. We would also like to point out that the earlier methods do not aim for automatic meeting minutes generation; instead, they strive to generate coherent meeting summaries in the form of paragraphs. Our motivation is to generate meeting minutes in the form of bullet points that adequately capture the contents of the meeting.

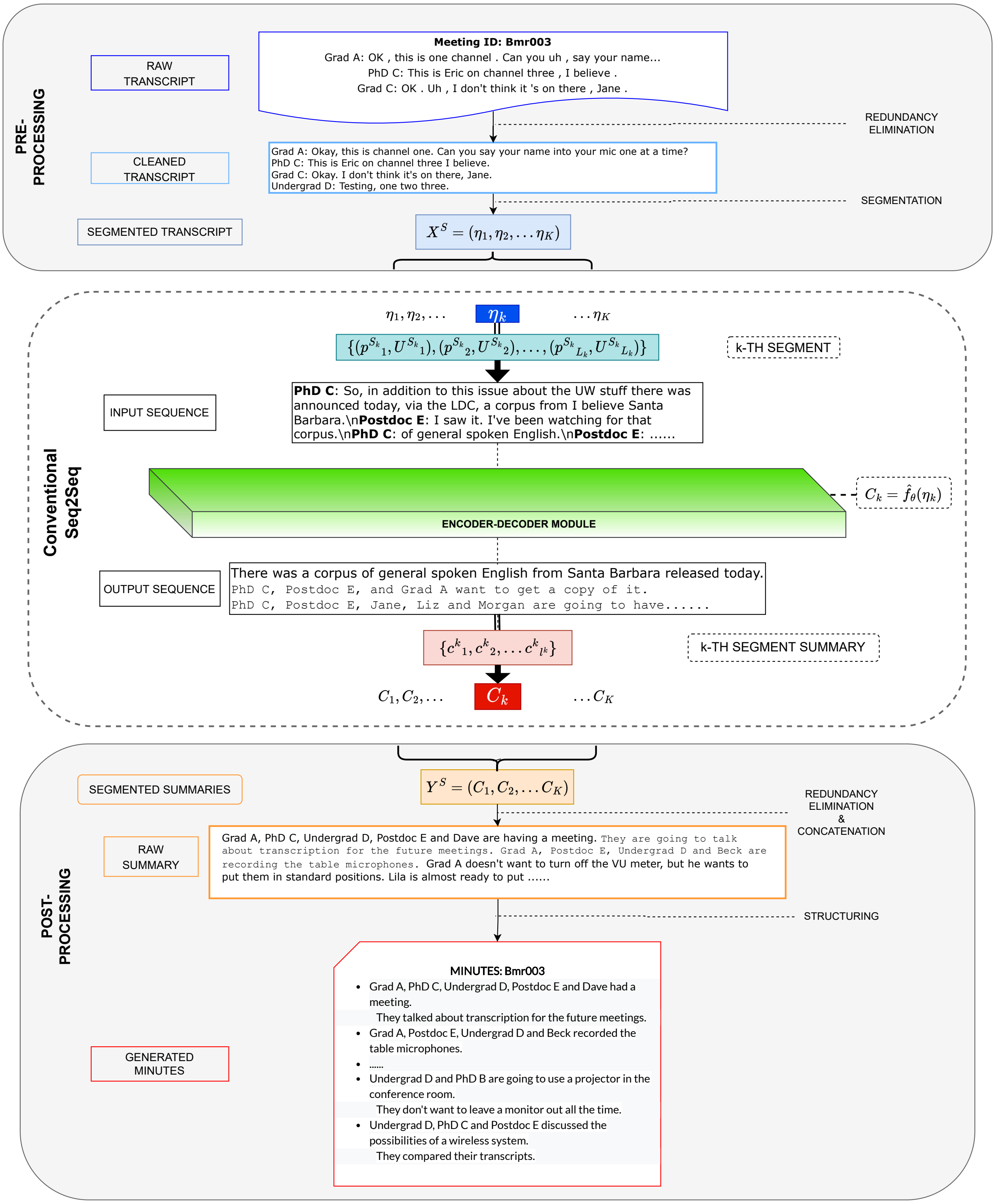
3 Methodology

Our current approach is inspired by one of the system submissions (Shinde et al., 2021) in the AutoMin shared task (Ghosal et al., 2021). Initially, we pre-process the transcripts as described in Section 3.1, later utilize the fine-tuned dialogue summarization model (Section 3.2), and finally, we post-process the outputs (Section 3.3). We describe the datasets used for the fine-tuning and evaluation in Section 4. We also provide automatic and human evaluation discussions and error analysis in Section 5. Kindly refer to Figure 1 for the entire system architecture.

3.1 Pre-Processing

Raw transcripts (directly from the ASRs) would require a good amount of pre-processing before one can proceed with the downstream tasks (automatic minuting in our case). In our experiments, the raw transcripts were already processed by human annotators to remove any inconsistencies during the respective corpora development. We discuss the steps employed for our use case on the already processed datasets.

Redundancy Elimination. Since current summarization models are not trained to eliminate redundancies and are often capped to specific input lengths, they struggle to process a long sequence of multi-speaker utterances and the dispersed information that comes with them (Ghosal et al., 2022b). We leverage specific pre-processing methods and em-



Meeting ID: Bmr003
Grad A: OK , this is one channel . Can you uh , say your name...
PhD C: This is Eric on channel three , I believe .
Grad C: OK . Uh , I don't think it 's on there , Jane .

Grad A: Okay, this is channel one. Can you say your name into your mic one at a time?
PhD C: This is Eric on channel three I believe.
Grad C: Okay. I don't think it's on there, Jane.
Undergrad D: Testing, one two three.

PhD C: So, in addition to this issue about the UW stuff there was announced today, via the LDC, a corpus from I believe Santa Barbara.\n**Postdoc E**: I saw it. I've been watching for that corpus.\n**PhD C**: of general spoken English.\n**Postdoc E**: ......

There was a corpus of general spoken English from Santa Barbara released today.
PhD C, Postdoc E, and Grad A want to get a copy of it.
PhD C, Postdoc E, Jane, Liz and Morgan are going to have......

Grad A, PhD C, Undergrad D, Postdoc E and Dave are having a meeting. They are going to talk about transcription for the future meetings. Grad A, Postdoc E, Undergrad D and Beck are recording the table microphones. Grad A doesn't want to turn off the VU meter, but he wants to put them in standard positions. Lila is almost ready to put ......

MINUTES: Bmr003

- Grad A, PhD C, Undergrad D, Postdoc E and Dave had a meeting.
  They talked about transcription for the future meetings.
- Grad A, Postdoc E, Undergrad D and Beck recorded the table microphones.
- ......
- Undergrad D and PhD B are going to use a projector in the conference room.
  They don't want to leave a monitor out all the time.
- Undergrad D, PhD C and Postdoc E discussed the possibilities of a wireless system.
  They compared their transcripts.

Figure 1: Architecture of the proposed *Automatic Minuting* pipeline.

ploy utterance cleaning and redundancy elimination based on thresholds to tackle this issue.

Consider a transcript with speaker-utterance pairs, $X^0 = \{(p^0_1, U^0_1), (p^0_2, U^0_2), ..., (p^0_L, U^0_L)\}$, where $p^0_j \subset P$, $1 \leq j \leq L$, is the j-th speaker and $U^0_j = (w^j_1, w^j_2 ..., w^j_{l_j})$ is the tokenized sequence of the j-th utterance; where $\{w^j_i\}$ represents the i-th token from the j-th utterance. For the j-th tokenized utterance, $U^0_j = (w^j_1, w^j_2 ..., w^j_{l_j})$ from the transcript, we generate a cleaned sequence, $U^c_j = (W^j_1, W^j_2 ..., W^j_{L_j})$, by eliminating repetitions, pauses and known special symbols for unarticulated sounds, unintelligibility, disfluency markers, and similar disruptions. We filter the utterances using custom stopwords set S that we define from various meeting transcripts from currently available corpora like AMI (McCowan et al., 2005), ICSI (Janin et al., 2003), and the dataset from AutoMin (Nedoluzhko et al., 2022). By this, we obtain the filtered utterance $U^f = (U^c \backslash S)$ and the corresponding context ratio R, which expresses how much the utterance was shortened by dropping stopwords compared to the cleaned version:

$$R = |U^f|/|U^c| \tag{1}$$

Ultimately, our processed transcript X' comprises utterances U^c_i where the ratio of non-stop-words R_i is big enough, i.e. $R_i \geq \alpha$ (α being a predefined threshold ratio).

Linear Segmentation. Current summarization models limit the length of input sequences they can process (Singh et al., 2021), so they cannot process the full-length transcripts in our data. Our approach here is simple: it breaks the transcripts into blocks with a uniform token length. We experiment with token lengths: 512, 768, and 1024, respectively.

Topical Segmentation. The linear segmentation technique is problematic whenever important information on a topic falls into the subsequent segment. To address this limitation, we experiment with two methods for topic-aware segmentation: Depth-Scoring (adopted from Solbiati et al. (2021)) and the TextTiling algorithm by (Hearst, 1993).

For Depth Scoring, we use a window of k_w segments, capping each segment to $\hat{L} = 60$ words and setting topic change threshold τ to 0.5 (these are tunable hyperparameters, kindly refer to Solbiati et al. (2021) for details). Let us consider Figure 2. For a transcript with N turns, we obtain their contextualized embeddings from an encoder. We apply max pooling on this embedding space.

For a pair of neighboring windows of segments, one consisting of turns $k - k_w$ till k, and the other of turns k till $k + k_w$, we obtain the cosine similarity, sim_k between the embeddings pooled across all segments in the respective windows. For a series of neighbouring window similarity scores $\hat{s} = (sim_{k_w}, ...sim_{N-k_W})$, we compute the depth scores as $dp_k = \frac{hl(k)+hr(k)-2sim_k}{2}$ where $hl(k)$ and $hr(k)$ are the highest similarity score on the left and right side of the k^{th} element in the series of similarity scores. We deduce the topic change indices with the help of the obtained window-similarity scores and depth scores. Following are the variations one can use while determining the topic change indices.

- **Segment-window capping.** With this approach, we compute the topic change indices as:

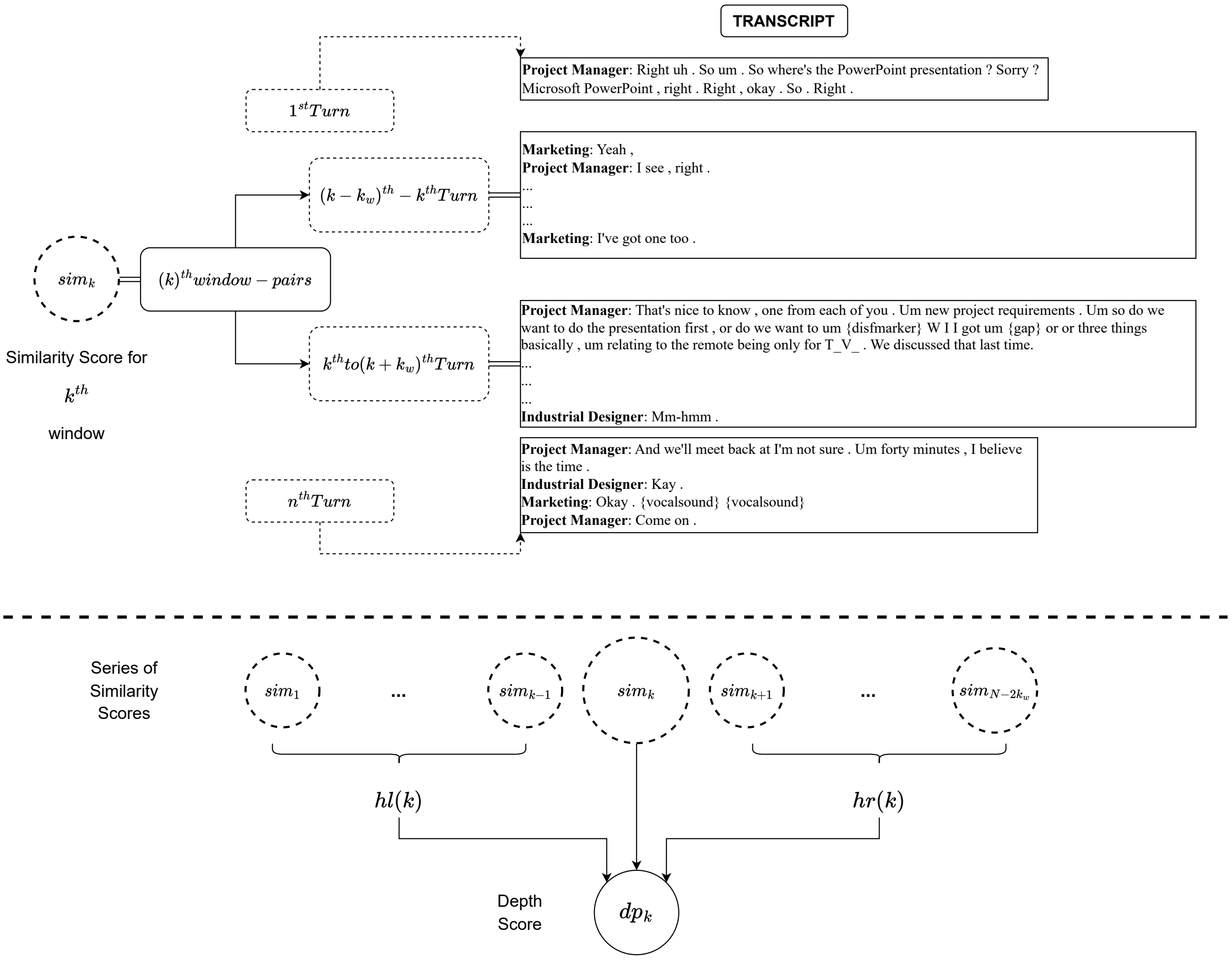
  $$T_{ds} = \{i \in [0, M] | sim_{k_w+i} \leq \mu_s - \sigma_s\} \tag{2}$$

  where T is the set of topic-change indices μ_s and σ_s are the mean and variance of the sequence, $M = N - k_W$ is the number of windows, sim_{k_w+i} is the similarity score of the i^th window.

- **TextTiling.** TextTiling is a method to subdivide texts into multi-paragraph units representing passages or subtopics by leveraging lexical co-occurrence and distribution patterns. Here, we use TextTiling to identify major subtopic shifts. After computing the window similarity scores, we use the TextTiling method to compute the segments in a transcript. For a series of depth scores $D = (d_1, d_2, ...d_{N-k_w})$, we compute the topic change indices as:

  $$T_{tt} = \{i \in [1, M] | d_i \geq \tau\} \tag{3}$$

Through one of the three approaches (linear, depth-scoring, or text tiling), we obtain the segmented transcript $X^S = (\eta_1, \eta_2, ...\eta_K)$ where $\eta_k = \{(p^{S_k}_1, U^{S_k}_1), (p^{S_k}_2, U^{S_k}_2), ..., (p^{S_k}_{L_k}, U^{S_k}_{L_k})\}$ is the sequence of speaker-utterance pairs belonging to that segment.



Project Manager: Right uh . So um . So where's the PowerPoint presentation ? Sorry ?
Microsoft PowerPoint , right . Right , okay . So . Right .

Marketing: Yeah ,
Project Manager: I see , right .
...
...
...
Marketing: I've got one too .

Project Manager: That's nice to know , one from each of you . Um new project requirements . Um so do we want to do the presentation first , or do we want to um {disfmarker} W I I got um {gap} or or three things basically , um relating to the remote being only for T_V_ . We discussed that last time.
...
...
...
Industrial Designer: Mm-hmm .

Project Manager: And we'll meet back at I'm not sure . Um forty minutes , I believe is the time .
Industrial Designer: Kay .
Marketing: Okay . {vocalsound} {vocalsound}
Project Manager: Come on .

Figure 2: Illustration of segment-windows, and calculation of similarity and depth scores referred in Section 3.1

We concatenate the speaker labels $p_{\bullet}$ with the corresponding utterances $U_{\bullet}$ and then across all items in the given segment back to the form of a single dialogue transcript. We then pass each of these plain text segments to one of the root summarization modules (see Section 3.2).

3.2 Summarization

We choose the pre-trained BART model (Lewis et al., 2019) in the summarization module in our pipeline. BART performs best among the other summarization models we tested, generating fluent and readable meeting minutes. Other summarization models include T5 (Raffel et al., 2019), Pegasus (Zhang et al., 2020), and RoBERTa2RoBERTa (Rothe et al., 2020). We fine-tune all these models on popular dialogue summarization datasets before integrating them into our pipeline.

BART is a denoising autoencoder for pretraining sequence-to-sequence models. The model is trained by corrupting text in an arbitrary noising function and then teaching it to reconstruct the original text. BART's ability to use source-side bi-directionality when operating on sequence generation tasks encourages its use for text summarization.

We pass the input sequence obtained from the preprocessing module through the summarization module. Again, for k-th segment, it returns a summary $C_k = \{c_1^k, c_2^k, ...c_{l_k}^k\}$, where c_i^k is the i-th summary line of the k-th segment. We rejoin all the segment summaries $Y^S = (C_1, C_2, ...C_K)$ to get the raw summary text.

Experimental Configuration We do not train any models from scratch but finetune most of them on

Datasets	# Dialogues	# Turns	# Speakers	# Turn Len.	# Len. of Dialogue	# Summary Len.	Compression
SAMSum	16.4K	11.2	2.4	9.1	124.0	23.4	81.12%
DialogSum	13.5K	9.5	2.0	15.8	168.5	25.8	84.70%
MediaSum	463.6K	30.0	6.5	49.6	1553.7	14.4	99.00%
AMI	137	535.6	4.0	10.4	5,570.4	321	94.24%
ICSI	59	819.0	6.3	10.5	8,567.7	576	93.28%
ELITR Corpus	124	254.4	5.8	9.7	8,890.8	387	95.65%

Table 1: Statistics of the dialogue and meeting summarization datasets we employ in our experiments. The top part is the larger summarization datasets we use for fine-tuning our models, the bottom part is theb meeting summarization datasets we use for model selection and testing. The reported statistics are averages across entire corpora. Lengths are in words. The compression ratio indicates how much the dialogue is shortened into the summary.

Datasets	Instances	Doc. Len.	Summ. Len.	% Comp.	%novel unigram
XSum	226.0K	488	27	94.5%	37.8%
CNN/DM	311.0K	906	63	93.0%	16.9%
R-TIFU	7.9K	641	65	89.9%	43.8%

Table 2: Document summarization datasets used for fine-tuning.

the data described in Section 4 below. For most models, a single Tesla K80 GPU is sufficient. Few larger models like BART-large and T5-large require multi-GPU training on NVIDIA GTX 1050 Ti or single GPU training on the NVIDIA A100-PCI-E-40GB variant. Training for individual finetuning procedures takes less than 2 hours, while warm-starting takes approximately 0.5 hours, depending on the dataset used. The hyperparameters and model configurations are consistent with the default values used during the pretraining of respective models. We set the finetuned BART on inference and generate our text with $num\_beams = 4, top\_k = 0.5$ and no limit on 'max_length'. We provide the hyperparameters and model configuration details in our code repository.

3.3 Post-Processing

After the main summarization, we use sentence compression methods, including swapping shortened phrases and pronouns and splitting longer sentences into two for improved readability. In our proposed pipeline, for each summary line, we filter out a set of unique entities (speaker names, project/corporation names, and location details). Further, we use a token-count threshold τ_{token} of 10 to include only those summary-sentences which are quantitatively informative enough (i.e., consisting of a minimum of τ_{token} number of tokens).

4 Dataset Description

Our work uses two types of data sources (see Table 1): (1) for fine-tuning summarization models, see Section 4.1, and (2) for the choice of the best setup and final evaluation of the minuting task, see Section 4.2.

4.1 Datasets for Fine-tuning Summarization Module

Here, we choose from some of the popular abstractive summarization datasets. Primarily, we use the dialogue summarization corpora: SAMSum (Gliwa et al., 2019), DialogSum (Chen et al., 2021), and MediaSum (Zhu et al., 2021).

Additionally, we use document summarization datasets XSum Narayan et al. (2018), CNN/DM Nallapati et al. (2016) and R-TIFU Kim et al. (2018), see Table 2. Their high compression ratio ("% Comp.") can potentially train the models to generate sequences more selectively, thus automatically eliminating redundancies.

4.2 Target Datasets: Automatic Minuting/Meeting Summarization

We primarily use ELITR Minuting Corpus (Nedoluzhko et al., 2022) for comparison with other systems. We further experiment on popular meeting summarization datasets: AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), due to similarities in the two tasks. AMI and ICSI come from staged product design meetings in companies, academic group meetings in schools, and similar arrangements. Each instance has a transcription of the entire dialogue and is annotated with a meeting summary and human-identified topic boundaries (except for ELITR Minuting Corpus). These meeting transcripts are extremely long, have a

	AMI			ICSI		
Models/Metrics	R-1	R-2	R-SU4	R-1	R-2	R-SU4
(A) Baselines and Comparing Systems						
Random	35.13	6.26	13.17	29.28	3.78	10.29
Cluster Rank (Garg et al., 2009)	35.14	6.46	13.35	27.64	3.68	9.77
Extractive Oracle	39.49	9.65	13.20	34.66	8.00	10.49
PGNet (See et al., 2017)	40.77	14.87	18.68	32.00	7.70	14.46
(B) Our best-performing setups						
bert2bert-cnndm-samsum	40.72	10.10	**27.13**	35.03	7.35	**24.48**
bart-xsum-dialogsum	42.40	10.34	17.67	36.95	6.94	13.68
t5-dialogsum	42.71	11.05	18.34	37.01	7.48	13.68
bart-xsum-samsum*	**45.17**	**13.30**	20.33	**38.75**	**8.51**	14.98
(C) State-of-the-art systems in Meeting Summarization						
HMNet (Zhu et al., 2020)	53.02	18.57	**24.85**	46.28	10.60	**19.12**
DialogLM (Zhong et al., 2021)	**53.70**	**19.60**	-	**49.50**	**12.50**	-
Summ^N (Zhang et al., 2021)	53.40	20.30	-	48.80	12.20	-

Table 3: ROUGE-1, ROUGE-2, ROUGE-SU4 scores of generated summaries on AMI and ICSI datasets. *→'bart-xsum-samsum' stands for our proposed model finetuned on the XSum corpus, further finetuned on the SAMSum corpus. Results in (C) are reproduced from the respective papers.

turn-based structure, and have multiple occurrences of redundant words and utterances.

Table 1 shows the relevant statistics of the dialogue and meeting summarization datasets that we use in our experiments.

5 Evaluation

In this section, we present the evaluation of our proposed pipeline in terms of automatic metrics in Section 5.1 and human evaluation metrics in Section 5.2. We compare our proposed pipeline with different summarization algorithms, finetune on combinations of abstractive summarization datasets, and report our performance on ELITR Minuting Corpus, AMI, and ICSI meeting summarization datasets.

5.1 Automatic Evaluation

For automatic evaluation, we make use of popular text summarization evaluation metrics. We report ROUGE (Lin, 2004) variants, namely ROUGE-1, ROUGE-2, ROUGE-SU4, which measures the overlap of unigrams, bigrams, and unigrams plus skip-bigrams (with max. skip of 4), respectively. We also provide METEOR (Banerjee and Lavie, 2005) scores which reward matching stems, synonyms, and paraphrases and not just exact matches.

5.2 Human Evaluation

To evaluate the quality of our output, we carry out a human evaluation of our minutes and compare it with the best-performing model outputs from the AutoMin 2021 shared task. Since we were the AutoMin shared task organizers, we had access to the human evaluators who also evaluated the system submissions in AutoMin. Six human evaluators rated our minutes in terms of *Adequacy, Grammaticality and Fluency* scores on a Likert scale of 5 (we report the average scores) (Ghosal et al., 2021). Because automatic metrics for text summarization evaluation have various shortcomings and are not apt to judge the quality of meeting minutes (Ghosal et al., 2022b), we attribute more importance to human evaluation, although the annotators were judging only our outputs in this run, without immediate comparison to AutoMin system outputs.

5.3 Results and Analysis

We discuss the experimental results and analyze the performance of our system in this section.

Table 3 compares the ROUGE scores of earlier models with our best setup (bart-XSum-samsum with linear segmentation). With no prior fine-tuning on AMI and ICSI meeting datasets, our pipeline outperforms several earlier approaches, including the popular Pointer Generator network (See et al., 2017) and the Extractive Oracle. However, the state-of-the-art models: HMNet (Zhu et al., 2020), DialogLM (Zhong et al., 2022) and SUMM-N (Zhang et al., 2021) are still superior in terms of the quantitative metrics.

Table 4 compares the automatic and human eval-

	Automatic Evaluation			Human Evaluation		
Model	R-1	R-2	R-L	Adequacy	Grammatical	Fluency
Ours-bart-xsum-samsum (Current Model)	**0.40±0.09**	**0.11±0.02**	0.18±0.03	**4.46/5.00**	**4.45/5.00**	**4.18/5.00**
Team ABC (Shinde et al., 2021)	0.33±0.08	0.08±0.04	**0.19±0.06**	3.98±0.73	4.45±0.37	4.27±0.55
Team Hitachi (Yamaguchi et al., 2021)	0.26±0.09	0.08±0.03	0.14±0.05	4.25±0.46	4.34±0.41	3.93±0.57

Table 4: Performance of our pipeline in comparison to the two best-performing participating systems at the **AutoMin Shared Task on the newly released ELITR Minuting Corpus.**

Model	**R-1**	**R-2**	**R-SU4**	**BERTScore**	**METEOR**
bart-xsum-samsum	**45.2**	**13.3**	**20.3**	**0.60**	**20.6**
bart-xsum-dialogsum	42.4	10.3	17.7	0.59	18.6
bart-base-samsum	39.9	11.2	16.1	**0.60**	15.1
bart-base-mediasum	33.2	7.0	11.3	0.55	14.0

Table 5: Comparison of the BART-based model setups with different finetuning datasets on the AMI test set.

Model	**Pk**	**WinDiff**	**ROUGE-1**	**METEOR**
Random	0.61	0.75	-	-
TextTiling	0.39	0.41	43.40	18.1
Capped	**0.34**	**0.35**	42.50	16.7
Linear (768)	0.44	0.50	**45.17**	**20.6**

Table 6: Comparison of the 'bart-xsum-samsum' model with different segmentation methods on AMI dataset

uation scores of AutoMin participating systems and our proposed model on ELITR Minuting Corpus (Nedoluzhko et al., 2022). Our model outperforms others on each of the metrics, confirming our pipeline's effectiveness. However, we must mention here that the scores of AutoMin systems were taken directly from Ghosal et al. (2021) and not remeasured in our annotation. The annotator pool was almost the same, and they had access to the AutoMin participant minutes. However, it is unlikely that they compared the AutoMin participants' outputs with the current system-generated output, so their evaluation scales could have shifted.

Table 5 shows the performance of our pipeline when used with different summarization models based on BART on the AMI test set. Our best-performing combination outscores the next by almost 3 points in terms of ROUGE-1; however, other model variants still perform close to the proposed approach. From the setups we tested, the best finetuning procedure starts with XSum and continues with the SAMSum dataset.

As we mentioned earlier, our model fine-tuned on the SAMSum corpus offers a better generation quality than those trained on other datasets. We attribute this to the fact that the dialogues in the SAMSum dataset are relatively simplistic and much more straightforward than those in DialogSum and MediaSum. The conversations are comparatively shorter and better reflect a conventional multiparty dialogue situation, leading to a better match between the training and testing conditions.

We also notice the differences caused by the training datasets used before the finetuning phase. Having a high compression ratio and novel word percentage, datasets like XSum demonstrate an extremely abstractive nature of summarization. Although the source text in XSum (Narayan et al., 2018) is longer than the dialogue instances from datasets like SAMSum (Gliwa et al., 2019), the summaries are relatively shorter. We observe a similar difference when we train the model on the XSum dataset compared to other datasets like the Reddit-TIFU (Kim et al., 2018) and the CNN/DailyMail (Nallapati et al., 2016). The generated minutes are relatively short, with more novel words and paraphrased sentences, qualities that are critical in abstractive summarization.

Furthermore, we carry out ablation experiments to check the effectiveness of our three segmentation methods (Table 6). We report the performance on the AMI dataset, which comes with reference segmentation of transcripts. We use the Pk (Beeferman et al., 1999) and WinDiff (Pevzner and Hearst, 2002) methods to evaluate the segmentation accuracy and report ROUGE-1 and METEOR scores on AMI transcripts.[1] We find ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) score surprisingly higher with linear segmentation (segments of at most 768 tokens). This result can be due to: with the topical segmentation schemes, several snippets of small talk and irrelevant information from the meeting make it to the final output, thereby affecting the overall scores.

[1] Ground truth segmentation of meeting transcripts are available for the AMI corpus.

Input: Transcript

Project Manager : Right uh . So um . So where's the PowerPoint presentation ? Sorry ? Microsoft PowerPoint , right . Right , okay . So . Right . Okay , so we've got uh so we've got new project requirements . Um . So basically we've got three things , and we've got forty minutes in which to uh {disfmarker} for this meeting to uh to discuss the various options . Um . Three presentations .
Industrial Designer : We have a {disfmarker} I guess we have a presentation each , 'cause I've got one . Um .
User Interface: Yeah .
Marketing : Yeah .
Project Manager : That's nice to know , one from each of you . Um new project requirements . Um so do we want to do the presentation first ,or do we want to um {disfmarker} W I I got um {gap} or or three things basically , um relating to the remote being only for T_V_ . We discussed that last time
. . .
Industrial Designer : Kay .
Marketing : Okay . {vocalsound} {vocalsound}

Output: Minutes

- Industrial Designer explains to Project Manager how the remote works.
  They have their own logo, color scheme and they have to incorporate it on the product.
  The current expansion of channels in the process of taking place, certainly the button up, down, but how many channels do we have to prepare for and how many buttons do we need?
- The Project Manager, Industrial Designer, Marketing and User Interface discussed the remote control functions.
- Industrial Designer will put delay to allow multiple numbers.
- Project Manager, User Interface, Marketing and Industrial Designer discussed the design of the remote control.
- Marketing, Industrial Designer, Project Manager and User Interface discussed the design of remotes for older televisions.
  There are 16 buttons on the front of the device: power button, standby, channel, volume, up, down, brightness, contrast, and a mute.
  On the back it will have the channel tuner, brightness and contrast.
- The project manager did the minutes of the meeting.

Figure 3: Sample output from our pipeline on a document from AMI corpus: Meeting Id-ES2014b

Figure 3 shows a sample minute generated from our pipeline approach. The transcript corresponds to 'ES2014b' from the AMI dataset. As we can see, the generated minute is coherent with the discussions from the meeting.

5.4 Error Analysis

We qualitatively examine and find that our outputs show the following categories of errors (Figure 4).

- **Made-up entities**. Anonymization of discrete entities in transcripts (e.g., LOCATION7, PERSON4, Marketing Manager) is consistent in most transcripts and minutes of our test datasets. Since no such anonymization is apparent in SAMSum, this sometimes results in the generation of made-up entities that are initially not part of that transcript.
- **Absence of context in summary**. Sometimes, the generated summary could use pronouns or other referring expressions from the transcript without ensuring that the element they are referring to is actually present in the summary. However, this issue is rare and did not occur in our final test runs.
- **Incomplete phrases**. Although less, we notice occurrences of incomplete sentences. These generally belong to those parts of the transcripts where the utterances either had missing punctuation or hesitations and interruptions on the speaker's part.

6 Conclusion

In this paper, we explore the use of large pre-trained language models fine-tuned on dialogue summarization datasets to automatically generate meeting minutes. We evaluate our proposed BART-based pipeline approach on the recently released corpus for automatic minuting (ELITR Minuting Corpus) as well as on the earlier AMI and ICSI meeting summarization corpora. We utilize existing multiparty meeting summarization datasets.

Our pipelined approach is promising and certainly puts up a case for further investigations to employ large language models for this challenging task. In future work, we would like to optimize our existing pipeline by replacing extractive filtering and utterance-level topic segmentation with an end-to-end method.

Acknowledgement

This work has received funding from the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

Case-1: Made-up entities

Instance - *“PhD A PhD F, PhD C and PhD F are discussing the encoding of things with time and data.”*

Explanation - It seems like as a normal summary line with correct grammar and readability. Consulting the transcript, we find out that ‘PhD C’ is not a real speaker, ’Grad C’ is the real speaker here. Hence, this is an error due to anonymization.

Instance - *“Marketing, Project Manager, Industrial Designer and Project Manager are meeting to....”*

Explanation - The ‘Project Manager’ was mentioned in the transcript once but it appears twice in the summary line. We attribute this to anonymization in the finetuning data, which collapses two people’s names into two very similar identifiers; the model then infers that repeating a similar (or even identical) entry is sometimes desired.

Case-2: Absence of context

Instance - *“PhD D discovered that on the wireless ones, you can tell if it’s picking up breath noises...”*

Explanation - The wording of the summary line uses a referring expression (‘the wireless ones’) without providing its referent in the surrounding lines.

Case-3: Incomplete phrases

Instance 1 - *“they don’t match well with the operating behavior of the — Marketing, Industrial Designer, Project Manager are discussing the design of the remote control”*

Instance 2 - *“They have decided to start with the black and white version. They will use double A or triple A batteries, rubberized buttons, a plastic casing for the plastic shell, a variety of designs, — Marketing Project Manager, Industrial Designer, User Interface and Project Manager are discussing the design of a keychain.”*

Explanation - Due to interruptions in the speech, the transcripts sometimes break one speech act into several utterance — often marked with a hyphen. This reflects in the model outputs as shown with a ‘—’ separator.

Figure 4: Error instances from the pipeline-generated summaries illustrating the error cases discussed in Section 5.4.

References

[Banerjee and Lavie2005] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

[Beeferman et al.1999] Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

[Celikyilmaz et al.2018] Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. *arXiv preprint arXiv:1803.10357*.

[Chen and Bansal2018] Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

[Chen and Metze2012] Yun-Nung Chen and Florian Metze. 2012. Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[Chen et al.2021] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

[Cho et al.2019] Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. *arXiv preprint arXiv:1906.00072*.

[Chopra et al.2016] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

[Fauville et al.2021] G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. 2021. Zoom Exhaustion & Fatigue Scale. *Computers in Human Behavior Reports*, 4:100119.

[Garg et al.2009] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. Clusterrank: a graph based method for meeting summarization. Technical report, Idiap.

[Ghosal et al.2021] Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021. Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–25.

[Ghosal et al.2022a] Tirthankar Ghosal, Marie Hledíková, Muskaan Singh, Anna Nedoluzhko, and Ondrej Bojar. 2022a. The second automatic minuting (automin) challenge: Generating and evaluating minutes from multi-party meetings. page TBA, july.

[Ghosal et al.2022b] Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2022b. Report on the SIGDial 2021 Special Session on Summarization of Dialogues and Multi-Party Meetings (SummDial). *SIGIR Forum*, 55(2), mar.

[Gliwa et al.2019] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

[Hearst1993] Marti A Hearst. 1993. Texttiling: A quantitative approach to discourse segmentation. Technical report, Citeseer.

[Janin et al.2003] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

[Jia et al.2020] Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631.

[Kim et al.2018] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv:1811.00783*.

[Lebanoff et al.2019] Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. *arXiv preprint arXiv:1906.00077*.

[Lewis et al.2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[Li et al.2019] Manling Li, Lingyu Zhang, Richard J Radke, and Heng Ji. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *57th Conference of the Association for Computational Linguistics*.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[Liu and Chen2019] Zhengyuan Liu and Nancy Chen. 2019. Reading turn by turn: Hierarchical attention architecture for spoken dialogue comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5460–5466.

[Liu and Chen2021] Zhengyuan Liu and Nancy F Chen. 2021. Dynamic sliding window for meeting summarization. *arXiv preprint arXiv:2108.13629*.

[Liu and Lapata2019] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

[Liu et al.2019] Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

[McCowan et al.2005] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

[Nallapati et al.2016] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

[Narayan et al.2018] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

[Nedoluzhko et al.2022] Anna Nedoluzhko, Muskaan Singh, Marie HledÃkovÃ¡, Tirthankar Ghosal, and OndÅ™ej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France, June. European Language Resources Association.

[Pevzner and Hearst2002] Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

[Raffel et al.2019] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

[Rothe et al.2020] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

[Rush et al.2015] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

[See et al.2017] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

[Shinde et al.2021] Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 26–33.

[Singh et al.2021] Muskaan Singh, Tirthankar Ghosal, and Ondrej Bojar. 2021. An empirical performance analysis of state-of-the-art summarization models for automatic minuting. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 50–60, Shanghai, China, 11. Association for Computational Lingustics.

[Solbiati et al.2021] Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. *arXiv preprint arXiv:2106.12978*.

[Wang et al.2020] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

[Xu and Durrett2019] Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. *arXiv preprint arXiv:1902.00863*.

[Xu et al.2019] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

[Yamaguchi et al.2021] Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, Ken ichi Yokote, and Kenji Nagamatsu. 2021. Team Hitachi @ AutoMin 2021: Reference-free Automatic Minuting Pipeline with Argument Structure Construction over Topic-based Summarization. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 41–48.

[Zhang et al.2020] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

[Zhang et al.2021] Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

[Zhao et al.2019] Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

[Zhong et al.2019] Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

[Zhong et al.2021] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *arXiv preprint arXiv:2109.02492*.

[Zhong et al.2022] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

[Zhu et al.2020] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv preprint arXiv:2004.02016*.

[Zhu et al.2021] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

Vanilla Recurrent Neural Networks for Interpretable Semantic Textual Similarity

Piotr Andruszkiewicz
Warsaw University of Technology
piotr.andruszkiewicz@pw.edu.pl

Barbara Rychalska
Warsaw University of Technology
b.rychalska@mini.pw.edu.pl

Abstract

Semantic similarity systems assess to what extent two words, phrases, sentences are similar in their meaning. In this task, rule-based and neural network systems achieve the best results. The former requires intensive human workload and the latter needs heavy computing. Can we achieve high accuracy without hand-crafted rules and without intensive computing? In this paper we present three types of Vanilla Recurrent Neural Networks that fulfill the aforementioned requirements for Interpretable Semantic Textual Similarity task and we compare them to the systems from SemEval competition.

1 Introduction

Semantic similarity, also called paraphrase detection, focuses on similarity of words, phrases, sentences in terms of semantic equivalence. Having a different grammar construction or different words, yet synonyms, used, two sentences may convey the same meaning. Semantic similarity assesses the level of semantic equivalence, e.g., by saying that a pair of sentences is not semantically equivalent or is equivalent to some extent.

There are many different specifications of semantic similarity task. We focus on two definitions of this task used in SemEval competition (Cer et al., 2015). The first one is a basic approach, called herein basic semantic similarity task, that expresses semantic similarity as one number from 0 to 5. 0 means that there is no semantic similarity, e.g., *A woman is slicing big pepper.* vs. *A dog is moving its mouth.* 5 is assigned when there is perfect semantic equivalence, e.g., *The man cut down a tree with an axe.* vs. *A man chops down a tree with an axe.* In the middle there are pairs that are similar to some extent, e.g., a pair *People are playing cricket.* vs. *Men are playing cricket.* is scored as 3.

A more complicated specification of a task in question, called Interpretable Semantic Textual Similarity (iSTS), defines 8 alignment types between chunks of sentences, e.g., EQUI - semantic equivalence – *in Olympics*, *at Olympics*, OPPO - semantic opposition – *lower* vs. *higher*. Additionally, similarity for each alignment is scored on a scale 0-5. In this paper, we focus on Interpretable Semantic Textual Similarity (iSTS) task as it gives more insight in the justification of the similarity.

We can distinguish two main types of systems that assess semantic similarity, namely; hand-crafted rule-based systems and statistical systems. The first group uses many rules prepared by linguists and is highly customized for a specific task. Such systems need a lot of manual work. Statistical systems use trained models to assess semantic similarity. Some of them also involve lots of manual work in features engineering. We focus on statistical systems that do not require manual work and here comes deep learning approach.

Moreover, we prefer network architectures powered by basic network cell, because amount of annotated data is limited for Interpretable Semantic Textual Similarity (iSTS).

Hence, we focus on simple network models; that is, Vanilla Recurrent Neural Network (Vanilla RNN). Vanilla networks have been successfully applied in various tasks, for instance, char, word sequences processing, including text generation (Karpathy and Fei-Fei, 2017) or handwriting generation (Graves, 2013). The network cell we use is much simpler and has less parameters to train than more complicated gates, for instance, GRU/LSTM. Thus, we present three new network architectures with basic network cell and compare them to a baseline architecture and other available solutions for Interpretable Semantic Textual Similarity with one-to-one relations. In our work, we focus on achieving high accuracy using vanilla Recurrent Neural Networks with basic network cell in Interpretable Semantic Textual Similarity (iSTS) with one-to-one relations. That has been achieved by refining network architectures and by using basic network cell.

2 Related Work

Semantic similarity systems could be divided into two main groups: rule-based systems and statistical systems. An example of a rule-based system is UMBC EBIQUITY (Han et al., 2013), which won SemEval 2013 (Diab et al., 2013). Despite the already mentioned drawbacks of rule-based approaches, such systems have recently dominated the area of semantic similarity. They use external resources prepared by linguists, e.g., databases of synonyms, and heavily depend on manual work in rules creation. The main idea behind this kind of systems is to find words that are semantically similar in both sentences and calculate the level of semantic similarity for each pair. Then, the aggregated similarity is calculated. These systems could be supported by statistical models, however, the influence of a statistical model on the whole system is usually rather low.

The second group - statistical systems - are often based on neural network models. The first neural network that achieved high accuracy in semantic similarity task was presented in (Socher et al., 2011). It was based on autoencoder that encodes a sentence and then decodes it into another one being as close to the original sentence as possible. In this approach two encoded sentences are compared and assigned a score. The encoder transforms a sentence according to a dependency tree. Recently, recurrent neural networks (RNN) have been used for semantic similarity task. In (Tai et al., 2015), LSTM gate organized in a tree structure architecture has been used. The tree was built according to a dependency parse tree. In (Mueller and Thyagarajan, 2016) siamese RNNs with tied weights between networks were used. We do not apply this restriction in our solution. RNN approach is the current trend in semantic similarity.

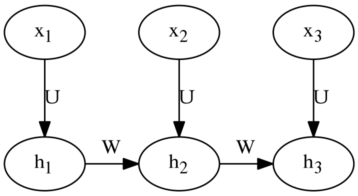


Figure 1: Linear Vanilla RNN.

There are also hybrid systems that combine rule-based approaches and statistical models. As they utilize advantages of these two kinds of systems, hybrid systems achieve high accuracy. A system of this type (Rychalska et al., 2016) won, for instance, semantic similarity task for English at SemEval 2016. Unfortunately, hybrid systems inherit also disadvantages of both of their combined approaches.

Our models are different from the above systems as they use statistical approach without time consuming manual work and apply basic network cell due to small available annotated data for iSTS.

3 Network Architectures

In this section, we present the baseline network architecture for Interpretable Semantic Textual Similarity task. We also propose three new architectures of recurrent neural networks built on top of basic network cell.

3.1 Baseline Linear Architecture

We use a linear vanilla architecture, denoted as VRNNH1-lin and shown in Figure 1, as a baseline for our three more complicated, yet still simple, vanilla architectures for semantic similarity assessment. x_i nodes represent terms from a sentence. Nodes may be a single value or a vector, e.g., an embedding vector (Turian et al., 2010), (Pennington et al., 2014) that represents semantics of a term. The hidden state of a network is calculated as $h_i = tanh(U \cdot x_i + W \cdot h_{i-1})$, where h_{i-1} is previ-

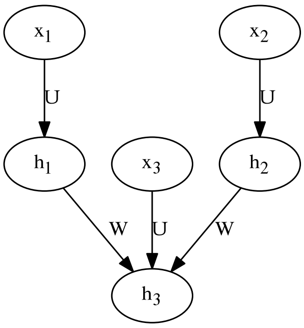


Figure 2: Tree Vanilla RNN.

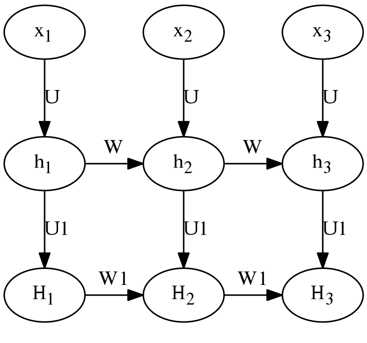


Figure 3: Linear Vanilla H2 RNN.

ous hidden state of the network and U, W are weight matrices.

3.2 Proposed architectures

The first of the proposed vanilla architectures (herein called VRNNH1-tree) is presented in Figure 2. A structure of a network is constructed according to a dependency tree of a sentence. The idea is similar to (Socher et al., 2011) (Tai et al., 2015), however, we use simpler vanilla unit. Compared to (Socher et al., 2011) we do not assume a decoding part of a network, as it is done in autoencoders. Moreover, we do not restrict a tree to be binary. Contrary to a linear network, a tree structured network does not process terms in the natural order as human beings do. However, we assume that a grammatical structure, represented by a parse tree, will be more beneficial for semantic representation of a sentence. The hidden state of a network is calculated as $h_i = tanh(U \cdot x_i + \sum_{j=1}^{m} W \cdot h_{i-1,j})$.

To extend the linear vanilla architecture, we add one more hidden layer. The extended architecture is depicted in Figure 3. A vector, H_i, for a node in the second hidden layer is computed as a non-linear function of a current hidden state h_i and a previous hidden state h_{i-1}. The additional non-linear function allows a network to model more complicated dependencies than those with only one hidden layer. Terms are processed according to a natural order of a sentence. h state is computed in the same way as in linear vanilla architecture: $h_i = tanh(U \cdot x_i + W \cdot h_{i-1})$. Additional hidden state is calculated as $H_i = tanh(U2 \cdot h_i + W2 \cdot H_{i-1})$. We denote this architecture as VRNNH2-lin.

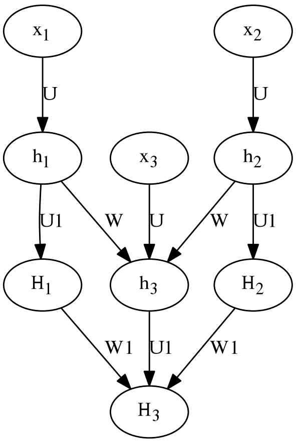


Figure 4: Tree Vanilla H2 RNN.

The same extension can be applied to VRNNH1-tree network, thus Figure 4 presents a tree based network with additional hidden layer called VRNNH2-tree. Compared to VRNNH1-lin we try to leverage grammatical structure of a sentence and allow for modeling of more complicated dependencies by incorporating the additional hidden layer. The hidden state in the first hidden layer of a network is calculated like in VRNNH1-tree, as $h_i = tanh(U \cdot x_i + \sum_{j=1}^{m} W \cdot h_{i-1,j})$. The state from the second layer is computed as $H_i = tanh(U2 \cdot h_i + \sum_{j=1}^{m} W2 \cdot H_{i-1,j})$.

In order to evaluate two chunks/sentences in the context of semantic similarity, we need one more layer which gives a final result. We chose a softmax layer (Su and Xu, 2015) for this purpose. The 6 output neurons represent one of the 6 possible results on a 0-5 scale for *basic semantic similarity* task and 8 neurons represent alignment type in *Interpretable Semantic Textual Similarity* task. Each neuron of a softmax layer gives a probability of a

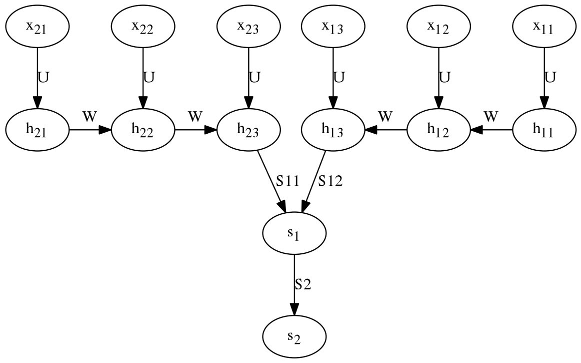


Figure 5: The full architecture of linear Vanilla RNN.

particular possible result. We can choose the result with the highest probability. A full network architecture combines final vectors representing two sentences and gives the final result. The full architecture for VRNN1-lin representation of a sentence is shown in Figure 5. h_{23} and h_{13} are nodes that output the vector representation for each of the two sentences. Then node s_1 compresses vectors representing sentences (e.g. 50 or 100 elements) to the length equal to the number of possible final outputs (e.g., six) for scoring estimation.

In the full network architecture presented in Figure 5, states for s_1 and s_2 are calculated as follows: $s_1 = tanh(S11 \cdot h_{23} + S12 \cdot h_{13})$, $s_2 = softmax(S2 \cdot s_1)$. For two hidden layer networks the final layers connect nodes from the second hidden layer, e.g., H_{23} and H_{13}. Please refer to Figure 6 that presents full network structure for VRNN2-lin sentence representation architecture. For tree structured architectures the idea of combining vectors representing each sentence is the same.

4 Non-Integer Scores

Scores provided in SemEval data are not always integers. Let us assume that the score is 3.75. Then we can guess that it comes from score 4 assigned by three annotators and score 3 assigned by one annotator. We need to choose the approach to process such scores in the softmax layer.

4.1 Scores Weighting

First approach is to reconstruct the distribution of scores. If we know that for 3.75 three persons assigned score 4 and one person chose score 3, we can represent a softmax layer as (0 0 0 0.25 0.75 0) where score 3 has a probability of 0.25 and score 4 has a probability of 0.75. Thus, we take into account minority votes.

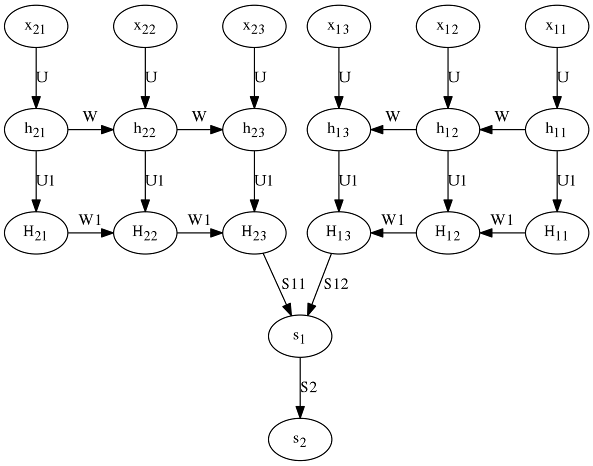


Figure 6: The full architecture of linear Vanilla H2 RNN.

Set	Training	Testing	Total
Images	375	375	750
Headlines	378	378	756

Table 1: Number of sentence pairs in the data sets.

4.2 Scores Rounding

We may also think, in case of score 3.75, that score 3 was a mistake or deviation from the proper score. Hence, we round the score to the most probable one and omit the minority votes by taking into account only majority votes. In such a case we would represent the softmax layer as (0 0 0 0 1 0) that means we assume the score of 4 should be assigned.

We examine both approaches in Section 5.

5 Experiments

We test the proposed architectures with Images and Headlines (Agirre et al., 2015) data sets from SemEval 2015 Interpretable Semantic Textual Similarity (iSTS) contest to compare our models to systems submitted by SemEval participants.The data sets contain around 750 sentence pairs each (please refer to Table 1. Golden chunks, scores and types are also provided.

We also use F measures applied in SemEval iSTS

Model	F1 score	F1 s+type	Avg
H1-lin	0.8620*	**0.6894**	**0.7757**
H1-tree	0.8516	0.6819*	0.7668*
H2-lin	0.8541	0.6679	0.7610
H2-tree	**0.8651**	0.6825*	0.7738*

Table 2: Experiments summary for 50 elements embedding vectors for Images set.

Model	F1 score	F1 s+type	Avg
H1-lin	0.8559*	0.6624*	0.7592*
H1-tree	0.8578*	0.6643*	**0.7611**
H2-lin	**0.8617**	0.6557*	0.7587*
H2-tree	0.8502	**0.6704**	0.7603*

Table 3: Experiments summary for 100 elements embedding vectors for Images set.

contest. *F1 score* (Agirre et al., 2015) is F1 measure that takes into account the score assigned to the alignment. The score should match. The alignment type is ignored.

F1 score+type (also *called F1 s+type*) (Agirre et al., 2015) is F1 measure that takes into account both alignment type and score for alignments.

In our tests we use golden chunks, thus the algorithm does not need to discover chunks within the sentences because it uses chunks provided in the data set.

In the experiments, we use word embeddings described in (Turian et al., 2010) and log-loss function as a loss function. To calculate a gradient, Limited-memory BFGS algorithm is used (Liu and Nocedal, 1989). Regularization is performed with L2 method. The source code is available at https://www.dropbox.com/s/ot25z73qnhue92f/vrnn1.0.zip?dl=0 (anonymized link).

Model	F1 score	F1 s+type	Avg
H1-lin	0.8560	0.6942*	0.7751*
H1-tree	0.8536	0.6953*	0.7745*
H2-lin	**0.8652**	0.6958*	**0.7805**
H2-tree	0.8509	**0.6981**	0.7745*

Table 4: Experiments summary for 50 elements embedding vectors and rounding for Images set.

5.1 Experiments with Short Word Embeddings

First we conduct experiments using word embeddings consisting of 50 elements. The results for Images set are shown in Table 2. The first column contains the names of network architectures. "H1" indicates a network with one hidden layer. "H2" means that two hidden layers are used. "lin" denotes a linear architecture. "tree" indicates that a network constructed according to the dependency tree is used. Thus, "H2-lin" denotes a network with linear structure and two hidden layers. The second and third column contains results of F1 score and F1 score+type, respectively. The last column presents the average of F1 score and F1 score+type measures. The best results for each measure or average is marked in bold. * denotes values for which the difference between them and the best value, marked in bold, is not statistically significant at 0.05 p-value level. We use weighting for non-integer scores (please refer to Section 4).

In terms of F1 score measure, the best result (0.8651) is obtained with the most complicated network, i.e., the tree structured network with two hidden layers. However, the linear network with one hidden layer obtains the result that is not statistically different from the one marked in bold. For F1 score+type the best result (0.6894) is achieved by the simplest architecture, i.e., the linear network with one hidden layer. The same network gets the best average result (0.7757).

Comparing the statistically differences, only the linear network with two hidden layers yields the lowest results obtaining statistically lower results three times.

Table 6 presents the results for Headlines set. Considering statistic significance, the presented output of four networks does not differ for all measures, thus, all four networks perform well.

5.2 Experiments with Longer Word Embeddings

In the next set of experiments, we use longer word embedding vectors with 100 elements. Table 3 summarizes the results for Images set. In this setup, the best result for F1 score is obtained by the linear network with two hidden layers (0.8617). However, only H2-tree network yields statistically different re-

Model	F1 score	F1 s+type	Avg
H2-lin	**0.8652**	0.6958	**0.7805**
H2-tree	0.8509	**0.6981**	0.7745
NeRoSimR1 (Banjade et al., 2015)	0.7877	0.5841	0.6859
UMDuluthBlueTeam2 (Karumuri et al., 2015)	0.7968	0.5964	0.6966
FULL (Lopez-Gazpio et al., 2017)	0.8085	0.6159	0.7122

Table 5: Experiments summary for 50 elements embedding vectors and rounding for Images set.

Model	F1 score	F1 s+type	Avg
H1-lin	0.8632*	0.6615*	0.7623*
H1-tree	0.8648*	**0.6708**	**0.7678**
H2-lin	0.8662*	0.6651*	0.7657*
H2-tree	**0.8675**	0.6598*	0.7637*

Table 6: Experiments summary for 50 elements embedding vectors and Headlines set.

Model	F1 score	F1 s+type	Avg
H1-lin	0.8680*	0.6557*	0.7619*
H1-tree	**0.8700**	**0.6641**	**0.7670**
H2-lin	0.8679*	0.6543*	0.7611*
H2-tree	0.8679*	0.6514*	0.7597*

Table 7: Experiments summary for 100 elements embedding vectors and Headlines set.

sults from the one marked in bold.

For F1 score+type the network with two hidden layers and tree structure performs the best and achieves 0.6704. The average points also tree structured network but with one layer that yields 0.7611. The differences between results of different models for F1 score+type and average are in the range of around 1–1.5 percentage points (p.p.) and are not statistically different.

For Headlines (Table 7 once again, we obtain the results which are not statistically different.

Compared to 50 elements word embedding vectors and Images set, the networks trained with longer vectors perform better only for H1-tree and H2-lin in terms of F1 score measure. All averages are lower than results obtained for 50 elements word embedding vectors and Images set. For Headlines, the differences between averages for 50 and 100 elements vectors are so small that they are not statistically significant.

Model	F1 score	F1 s+type	Avg
H1-lin	**0.8734**	0.6783*	0.7758*
H1-tree	0.8680*	0.6702	0.7691
H2-lin	0.8694*	0.6747*	0.7720*
H2-tree	0.8707*	**0.6830**	**0.7769**

Table 8: Experiments summary for 50 elements embedding vectors and rounding for Headlines.

Thus, not only do longer vectors not improve the models but even lower the results, especially for F1 score+type measure. Hence, for further experiments we choose 50 elements word embedding vectors.

5.3 Non-Integer Scores

We also test the alternative approach to non-integer scores; that is, rounding (for details please refer to Section 4). The results are presented in Table 4. "H2-lin" achieves the best results for F1 score obtaining 0.8652. For F1 score+type and average the networks obtain the results that are not statistically different. These are the highest values for all set of experiments and measures for Images data set. Rounding increases especially F1 score+type measure, which is above 0.69 for all network architectures.

For Headlines data set (please refer to Table 8) only H1-tree is significantly worse than other networks for F1 score+type and average. Compared to previous results, this set of experiments obtains the best results for Headlines also.

The rounding approach resembles majority voting and reduces non-agreement between annotators. For scores close to integers, e.g., 4.75, one score of 4 may be a mistake or deviation from major score and it is reduced by rounding the value to 5. This kind of approach suggests the network the strict answer and reduces uncertainty about the golden score.

Model	F1 score	F1 s+type	Avg
H1-lin	**0.8734**	0.6783	0.7758
H2-tree	0.8707	**0.6830**	**0.7769**
NeRoSimR3 (Banjade et al., 2015)	0.8157	0.6426	0.7326
NeRoSimR2 (Banjade et al., 2015)	0.8263	0.6401	0.7332
FULL (Lopez-Gazpio et al., 2017)	0.8211	0.6185	0.7198

Table 9: Experiments summary for 50 elements embedding vectors and rounding for Headlines set.

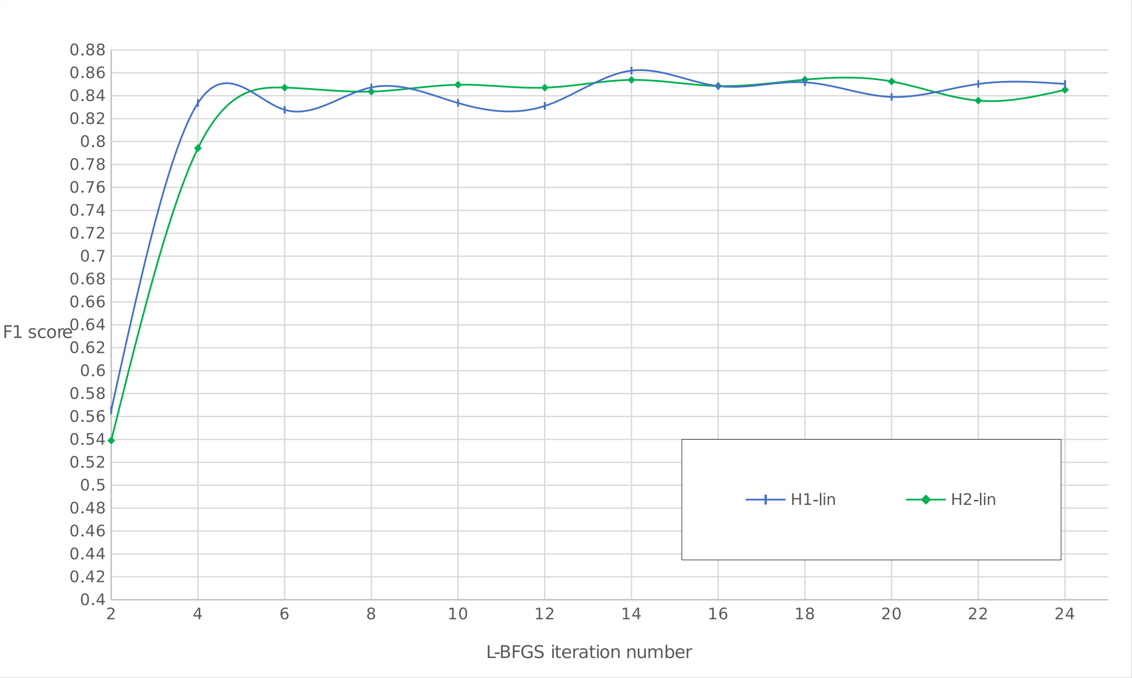


Figure 7: F1 score measure of linear Vanilla RNN with one and two hidden layers on Images set with respect to the number of iterations.

To sum up the experiments, our models assess semantic similarity quite well in iSTS. In this task, chunks are scored. Chunks usually are shorter, contain less tokens, than whole sentence, thus vanilla network with basic network cell may process it quite well as the relations within chunks are simpler than in whole sentences. This also could be the reason for networks based on parse tree not obtaining significantly better results than linear architectures. As the relations in chunks are simpler than in whole sentences additional information provided by parse tree does not help the system. Moreover, Vanilla Recurrent Neural Networks perform well in iSTS on small amount of data.

5.4 Comparison with Other iSTS Systems

As the networks trained on 50 elements word embedding vectors and rounding achieve the best results, we choose our best models to represent our approach.

We compare our models to systems from SemEval 2015 (Agirre et al., 2015) as there were changes introduced in SemEval 2016 and our models need to be adjusted. In SemEval 2015 assumed one-to-one alignments and SemEval 2016 introduced many-to-many alignments. We plan to do it in future work. We cannot compare results of systems prepared for basic semantic similarity task either as iSTS task is more complicated and systems designed for basic semantic similarity task cannot be used for iSTS task without further substantial modifications. The results of the comparison are shown in Table 5 and 9.

The system that obtained the best result for Images in SemEval 2015 iSTS for F1 score (0.7968) and F1 score+type (0.5964) measures was UMDuluth BlueTeam (Karumuri et al., 2015). NeRoSim (Banjade et al., 2015) ranked II for Images set, as it achieved 0.7877 for F1 score and 0.5964 for F1 score+type. System FULL (Lopez-Gazpio et al., 2017) outperforms two aforementioned systems. It is the solution prepared by a team that relates to organizers of SemEval 2015, thus the system could not be ranked. It scored 0.7122 in average measure.

Both our models outperform other systems in all measures for Images data set. Even all our remaining models achieve better results in terms of F1 score, F1 score+type, and average. The difference between our best model and the other best system is around 6 p.p. in F1 score, 8 p.p. in F1 score+type, and 7 p.p. in average, which is high improvement.

For Headlines our best models shown in Table 9 outperform significantly other systems also. The difference between our best model and the other best system is around 5 p.p. in F1 score, 3 p.p. in F1 score+type, and 4 p.p. in average. Moreover, our remaining models also yields better results the other systems.

The mentioned systems from SemEval 2015 used extensive feature engineering. In Interpretable Semantic Textual Similarity, NeRoSim applied hand-crafted rules. The system UMDuluth BlueTeam was a hybrid of human tuned word aligner, supervised machine learning and even translation systems. Our system achieves better results even though it does not use feature engineering nor customization performed by humans.

5.5 Iterations

Figure 7 presents the example F1 score measure with respect to number of L-BFGS algorithm iterations. We show linear networks with one and two hidden layers. 50 elements word embedding vectors are used and weighting for non-integer scores. Our models usually achieve high accuracy between 6th-8th iteration and remain stable for further iterations. The models with two hidden layers usually still yield lower F1 than models with one hidden layer at 4th iteration.

The basic network cell makes the process of training a model effective despite small number of available samples. It proves its usability in Interpretable Semantic Textual Similarity task.

6 Conclusions and Future Work

We presented three new architectures of recurrent neural networks built on top of basic network cell. Our aim was to maintain or increase the accuracy of established models by using Vanilla Recurrent Neural Networks and basic network cell. This approach was motivated by small amount of data available for Interpretable Semantic Textual Similarity (iSTS).

The proposed models are promising in iSTS as they achieved better results compared to the heavily hand-crafted systems. And there is still a lot of room for improvements in our models. We could also combine them with other hand-crated systems to obtain better results. However, we would like to stay away from feature engineering as much as we can, since this process is not automatic and requires manual work. In future work, we plan to check the accuracy of more complicated gates, such as GRU or LSTM, and test their influence on both accuracy and training time. Moreover, we would like to apply siamese networks to iSTS. We would also like to propose different network architectures tuned for Interpretable Semantic Textual Similarity task. We plan to adjust our system to modified iSTS SemEval 2016 task by introducing many-to-many alignments and test its accuracy.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Cer et al. (Cer et al., 2015), pages 252–263.

Rajendra Banjade, Nobal Bikram Niraula, Nabin Maharjan, Vasile Rus, Dan Stefanescu, Mihai C. Lintean, and Dipesh Gautam. 2015. Nerosim: A system for measuring and interpreting semantic textual similarity. In Cer et al. (Cer et al., 2015), pages 164–171.

Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. 2015. *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*. The Association for Computer Linguistics.

Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors. 2013. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*. Association for Computational Linguistics.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. Umbc_ebiquity-core: Semantic textual similarity systems. In Diab et al. (Diab et al., 2013), pages 44–52.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Sakethram Karumuri, Viswanadh Kumar Reddy Vuggumudi, and Sai Charan Raj Chitirala. 2015. Umduluth-blueteam: SVCSTS - A multilingual and chunk level semantic similarity system. In Cer et al. (Cer et al., 2015), pages 107–110.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528.

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowl.-Based Syst.*, 119:186–199.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2786–2792. AAAI Press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP team at semeval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 602–608. The Association for Computer Linguistics.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 801–809.

Hang Su and Haihua Xu. 2015. Multi-softmax deep neural network for semi-supervised training. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dres-*

den, Germany, September 6-10, 2015, pages 3239–3243. ISCA.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566. The Association for Computer Linguistics.

Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 384–394. The Association for Computer Linguistics.

Filipino Undergraduates' Perceptions of Translanguaging in a Linguistically Diverse Context

King Arman A. Calingasan
De La Salle University
Taft Avenue
Malate, Metro Manila
king.calingasan@dlsu.edu.ph

Earl Emunahlyn Joice V. Erlano
Notre Dame University
Notre Dame Avenue
Cotabato City, Maguindanao
erlano.eej@gmail.com

Haina P. Salik
Notre Dame University
Notre Dame Avenue
Cotabato City, Maguindanao
salikhaina6@gmail.com

Abdullaken U. Sinagandal
Notre Dame University
Notre Dame Avenue
Cotabato City, Maguindanao
sinagandal.laken@gmail.com

Abstract

This explanatory sequential mixed-method study aimed to determine the general perception of undergraduate students toward translanguaging and the reasons behind these perspectives. This study was done in two phases: a cross-sectional quantitative approach and a qualitative semi-structured interview. The data collected from the two phases were used to explain and interpret the perceptions of the participants (N=1170). It was conducted in a private higher education institution in Central Mindanao where undergraduates were mostly bilingual or multilingual. Overall, based on the descriptive statistics analyses, the study found that the majority positively perceived translanguaging as a practice, for second language learning, in social settings, and in higher education. Moreover, using thematic analysis, the study discovered the major reasons why college students had positive perceptions of translanguaging. They perceived that it (a) is beneficial in classroom activities, (b) is an effective tool to communicate and express ideas, and (c) allows effective learning. Therefore, the present study recommends the integration of translanguaging in the teaching and learning processes.

1 Introduction

In recent years, several researchers have become interested in investigating the various stakeholders' perceptions of translanguaging (Daniel & Pacheco 2016; Moody et al. 2019; Torpsten, 2018; Yuvayapan, 2019). Fang and Liu (2020) commented that studying translanguaging will give a broader view of it in classroom discourse practice. One advantage of studying perceptions about translanguaging is that the results can help stakeholders figure out whether or not it is a good thing to practice in the teaching-learning process. Therefore, it is beneficial to outline different perspectives and examine how it affects learning. This demand of knowing the perspectives about the practice urges the researchers to conduct a further study that aims to understand different views from another category of participants.

Some teachers in bilingual or multilingual classrooms across the globe usually practice translanguaging. These teachers are sometimes unaware that they apply translanguaging in the teaching process and sometimes deny the use of this approach. It can also occur naturally and often unexpectedly. This shows that monolingual educational policies cannot control translanguaging practices (Canagarajah, 2011). While there are existing misconceptions of translanguaging such as interpreting it as code-switching, it is broader and more likely a practice that allows speakers to use more than one language in a systematic and deliberate way to elicit learning. Translanguaging, as defined by Moody et al. (2019), is a 21st-century approach to language, bilingual education, and bilingualism that uses all linguistic resources of learners without separating their use and is critical to the student learning experience. On the contrary, Tabatabaei (2019) emphasized that the effects of translanguaging in learning are not yet entirely determined as positive or negative. However, she also claimed that it is much more functional nowadays.

In the earlier studies, researchers have found that graduate students (Moody et al. 2019), high school students (Daniel & Pacheco 2016), elementary students (Torpsten, 2018), and teachers (Yuvayapan, 2019) have contrasting views regarding the use and practice of translanguaging. For instance, the findings in the study of Moody et al. (2019) showed that graduate students had a positive perception of translanguaging and hence considered it a great practice, whereas neutral perspectives among teachers and students were found in the study of Fang and Liu (2020). In addition, Rivera and Mazak (2017) revealed that students held different beliefs about the use of translanguaging inside the classroom setting. Some students considered translanguaging useful in times of high-stress situations, while other participants complained that the use of multiple languages in classroom instruction made it difficult for them to retain the original information. Furthermore, Yuvayapan (2019) observed that translanguaging practices were hindered because of the expectations coming from the stakeholders, that is, monolingualism is the only approach that should be used in teaching. The said expectations, however, were contradicted by Galante (2020) who stressed the need for teachers to be familiar with the approach of translanguaging.

Additionally, Wang (2016) noted that teachers' and students' views and behaviors about translanguaging are helpful in scaffolding approaches that will improve the students' engagement and their relationship with teachers. In recent existing studies, researchers have suggested exploring the specific reasons of students why they have certain perceptions regarding the use of translanguaging (Fang & Liu, 2020; Moody et al., 2019; Rivera & Mazak, 2017). Most of them are quantitative, such as the study of Moody et al. (2019) and Khairunnisa and Lukmana (2020) which fail to explain the reasons for the positive and negative views on the use of translanguaging; and qualitative, such as the study of Galante (2020), and Yuan and Yang (2020) which lack numeric data. The separate use of these two different research designs creates possible gaps that may need to be filled in. Therefore, it is essential to note that a mixed-method design will be much more appropriate to utilize, for it will provide a quantitative interpretation as well as a qualitative explanation in the study. Along with that, other studies vary in terms of their participants. Particularly, graduate students were the chosen participants in the study of Moody et al. (2019) because undergraduates from their research setting were monolingual and might not have sufficient experience regarding the use of translanguaging.

To fill these gaps, the present study attempts to explicate Filipino undergraduates' perceptions of translanguaging in higher-level education. Specifically, it seeks to determine how these bilingual and multilingual undergraduates view translanguaging as a practice, for L2 learning, in social settings, and in higher education. Moreover, because most studies on perceptions of translanguaging are conducted quantitatively and qualitatively, this study uses a mixed method to incorporate a quantitative and qualitative phase to provide a general interpretation and an in-depth understanding of the students' reasons behind their specific perceptions. Lastly, the findings of this study provide implications for English language education and problematize the hegemonic role of English as the primary medium of instruction for all subjects except Filipino courses at the tertiary level in the Philippines.

1.1 The Present Study

Previous studies revealed that there are varying perceptions about translanguaging across different categories of participants, yet only a few researchers considered studying the undergraduates' perceptions of translanguaging. Therefore, the present study attempts to investigate the perceptions of undergraduate students toward translanguaging practices in a linguistically diverse classroom. Furthermore, the researchers invited a large number of Filipino bilingual and multilingual undergraduates to participate in the study compared with the studies previously conducted. Most importantly, this current research employed a mixed-method approach to further examine the reasons behind the students' perceptions of translanguaging. It specifically seeks to answer the following research questions:

1. How do Filipino undergraduate students perceive translanguaging as a.) practice, b.) for second language learning, c.) in social settings, and d.) in higher education?
2. What are the reasons why students perceive translanguaging positively?

2 Methodology

2.1 Research Design

This study utilized an explanatory sequential mixed-method approach, which is often used in collecting data during a particular period of time in two subsequent steps (Ivankova et al., 2006). This approach included the use of quantitative and qualitative design, respectively. Edmonds and Kennedy (2016) explained that an explanatory sequential design is an approach used when a researcher wants to compare quantitative and qualitative data. In the present study, qualitative data were used in interpreting and clarifying the quantitative data analysis results. Moreover, this approach was beneficial to the study because the student's responses to the quantitative survey questionnaire were explicated in their interview responses. In other words, the qualitative data expounded the tabulated and analyzed data.

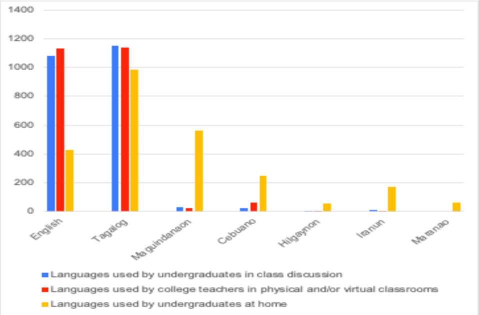
In the quantitative phase, a cross-sectional design was employed in which researchers measured the result as well as the exposure of the respondents at a single point in time (Setia, 2016). It was appropriate for this study because cross-sectional data are efficient when used in descriptive types of studies that are assumed to be analytical (Zangirolami-Raimundo et al., 2018). Moreover, this method helped the study to provide numerical data vis-à-vis respondents' perceptions of translanguaging.

To further explain the results of the quantitative data, this study employed a qualitative interview among the purposely selected participants. Moreover, it was well-suited and beneficial because it responded to the hows and whys of the study, rather than how many or how much (Tenny et al., 2017), and gave a broader explanation of the reasons for the respondents' perceptions.

2.2 Research Setting

The current study was conducted at a private university in Cotabato City, Central Mindanao, Philippines. This city is inhabited by several ethnolinguistic groups, where the most widely spoken Philippine languages are Maguindanaon, Tagalog, Cebuano, Hiligaynon, Maranao, and Iranun (Philippine Statistics Authority [PSA], 2013). Moreover, the majority of the students enrolled in this university are bilingual and multilingual, making it the most appropriate setting for the current study.

Figure 1. Languages Used at the University and at Home



Although the participating school implemented a monolingual policy in accordance with the Executive Order 210 series 2003, it is safe to assume that the use of translanguaging is still common in the learning context of these students. As Canagarajah (2011) argued, monolingual educational policies cannot control translanguaging practices *in a linguistically diverse learning environment* [emphasis added]. In fact, in this study, the majority of the participants reported that they speak English and Tagalog in class discussions, and at the same time, their teachers used these two languages in physical and virtual classrooms (see Figure 1). It is also interesting to note that Maguindanaon and Cebuano languages were seldom used by both students and teachers.

2.3 Research Participants

The participants of this study were bilingual and multilingual undergraduate students coming from different departments in a private institution. They were selected because a few studies on perceptions of translanguaging have considered this group of stakeholders.

In the quantitative phase of the study, 1170 (41.29%) out of 2,833 enrolled undergraduates responded to the web-based survey. Seventy percent were female, while 28.1% were male. The remaining 1.9% preferred not to say their biological sex. In addition, the respondents were from different year levels (n= 404 first-year students; 308 second-year students; 317 third-year students; 140 fourth-year students; 1 fifth-year student) with ages ranging from 18 to 28. Lastly, their home languages were Tagalog (84.3%), Maguindanaon (48%), English (36.5%), Cebuano (21%), Iranun (14.4%), Maranao (5.2%), Hiligaynon (4.7%), other local languages of the minorities (0.17%), and foreign languages like Arabic, Bahasa Melayu, and Mandarin (0.5%). It is important to indicate that there were students who reported that they have more than one home language.

Based on the results from the quantitative data analysis, participants in the qualitative phase were purposely selected following the set criteria: (a) they responded to the quantitative survey questionnaire, and (b) they had the highest (4.00) mean score in the survey, which means they viewed translanguaging positively. The researchers chose two participants from each college who had a positive perception of translanguaging based on the descriptive analysis of the quantitative data. Therefore, a total of 10 participants were invited to participate in the semi-structured interview to further share the reasons behind their views on translanguaging. Only students with positive perceptions of translanguaging were chosen because the quantitative analysis of this study provided findings that the majority of the respondents positively viewed translanguaging.

2.4 Research Instruments

Survey Questionnaire. The study utilized a survey questionnaire to get the desired information from the respondents about their perceptions of translanguaging. The items in this questionnaire were originally developed by Rivera and Mazak (2017) and were modified by Moody et al. (2019) in their study. The Likert scale that originally consisted of five-point items was reduced to four in the present study, removing the neutral option. This made it consist of items one to four, with one as strongly disagree and four as strongly agree. The reason for deleting the neutral option is the difficulty of providing accurate interpretations of the neutral responses. Although some argue that it is much better to have longer scales to test consistency, Blasius and Thiessen (2001) claimed that neutral categories (e.g., neither disagree nor agree) may be treated as hidden nonresponses that implied respondents' noncommittal opinions. Furthermore, items 1, 4, 5, 6, 8, 11, 19, and 20 in the survey questionnaire were reversely coded before the data analysis to form consistency. After pre-testing the instrument, this adapted survey was found highly reliable for the present research because it had an internal consistency of 0.927.

The content of the questionnaire which was created through an electronic form was divided into 6 sections. The first section stated the title and the purpose of the study. The second section consisted of five items that were designed to indicate the general information of the respondents. The participants responded according to their online synchronous and past face-to-face learning experiences. Next, the third section provided a simple definition of translanguaging and two situational examples of how translanguaging is applied by teachers in physical or virtual classes. To ensure that respondents understood the concept of translanguaging before answering the items in the survey, the fourth section asked respondents whether they comprehended it or not. If they did not grasp its idea, they were requested to discontinue answering the web-based survey. The fifth section consisted of 23 statements adapted from Moody et al. (2019) which were designed to explore the perceptions of students towards translanguaging. Lastly, the sixth section requested respondents to grant the researchers full consent to collect, store, access, and/or process their data whether manually or electronically, for the purpose and period allowed under the Republic Act 10173 otherwise known as the Data Privacy Act of 2012, and other applicable laws and regulations.

Semi-structured Interview. The study used a semi-structured interview guide for participants to explain their perspectives on translanguaging. Four interview questions were asked in the qualitative phase of the study. Codó (2008) considered an interview a useful method for gathering information on multilingualism. It includes personal and interpretive information and ideas, benefits, and impressions of their language and other language users' behaviors. Additionally, the interview comprised open-ended questions that allowed Filipino undergraduates to expound their views on translanguaging. These questions which were anchored on the quantitative result, specifically in the aspects of L2 learning and higher education, were checked and pre-tested to assess whether the questions captured the relevant information provided by the respondents. The semi-structured interview was an essential technique in this research because it collected and provided useful information that addressed the gaps in the previous studies.

2.5 Data Collection and Analysis Procedures

This study was conducted in two phases: quantitative and qualitative stages. After the study was approved by the research ethics committee, we conducted a pre-testing of the quantitative survey and the semi-structured interview. First, the adapted survey was administered to 29 students enrolled in the same university. There were no major issues found during the pre-testing after the respondents were debriefed about their experience in completing the survey. According to them, all items and even the instructions were clear and understandable. Most importantly, the instrument was highly reliable (α= 0.927).

Prior to the final data gathering, a letter of request with informed consent was sent to the respective deans of the five colleges. After getting the approval, we requested the official lists of students enrolled in the current semester. Then, we sent the online survey to the participants via email, disseminated the link to different group chats and Google classrooms, and shared it on Facebook. Although it was posted publicly online, the survey form was restricted only to a specific private university in Cotabato city and could only be accessed by using students' institutional email. Furthermore, they were informed that they could withdraw from the study during the process of collecting and analyzing the data.

Quantitative Phase. We collected data from the undergraduate respondents through a web-based survey which was encoded in the Google form. During the data gathering, we utilized a random sampling technique in choosing the participants. The duration of the data collection started from March 18 to 30 2022, garnering a total of 1186 respondents. However, during the deliberation of the responses, 16 students were considered unqualified respondents because they claimed that they did not understand the concept of translanguaging.

In the survey, students were provided with the definition of translanguaging, i.e., *As a pedagogical technique, translanguaging requires a more deliberate utilization of two languages in a teaching activity as opposed to just switching between them* (Yuvayapan, 2019). Two class situations where translanguaging is applied were also given after the definition (e.g., *The teacher allows the students to discuss the assigned topic of their group in their mother tongue. After the* *small group discussion, they are expected to present their topic to the class in English*). These 16 students who admitted that they did not understand the concept and examples of translanguaging continued answering the survey. Thus, their responses were removed from the raw data.

Furthermore, items 1, 4, 5, 6, 8, 11, 19, and 20 were reversely coded and then analyzed by getting the mean and standard deviation of the data. Afterward, the analyzed responses were used to answer the research questions.

Qualitative Phase. Based on the quantitative results, we drafted interview questions for the qualitative phase. We pre-tested the questions to two bona fide students from the same private institution. Similarly, no challenges were found; however, although the questions asked in English were comprehensible, interviewees preferred to code-switch from English to Tagalog and vice versa in responding to the questions. Mikuska (2017) asserted that pre-testing is vital in research to assess the viability and aptness of the interview questions before conducting the full-scale study.

After analyzing the quantitative data and pre-testing the interview questions, we purposely identified the participants for the semi-structured interview by determining those who got the highest (4.00) mean results. At least five of the participants in each college who viewed translanguaging positively were selected and contacted via email and Facebook messenger to participate in an interview through Google Meet.

The first two who responded became the participants in the qualitative phase and were individually invited to a virtual interview with the researchers. Before starting, each participant was presented with the interview protocol and was asked to sign the informed consent to allow the researchers to record the entire conversation. The interview questions guided the whole interview process.

After the virtual interview with each participant, the qualitative data were transcribed and thematically analyzed. Thematic analysis is an attainable, adaptable, and increasingly popular method for analyzing qualitative data (Braun & Clarke, 2012).

Participants were asked to share and expound their reasons behind their specific perceptions of translanguaging based on how they experienced it. Specifically, this research followed the procedures of Alase (2017). It started with the coding of qualitative data by reading the interview transcripts more than three times. In the coding process, we individually highlighted repeated keywords, phrases, or sentences from the participants' responses. During the coding, the participants were given anonymity codes (e.g., CED#1) to keep their identities confidential.

Before the analysis, we bracketed and suspended their personal perspectives of the phenomenon being studied so that their views will not influence the interpretations of the lived experiences of the participants. Next, exploratory comments on the highlighted statements were provided as annotations wherein related statements were grouped together (e.g., "*it helps us during brainstorming" and "I am using it when approaching my group members during the discussion"* can be grouped together as these are *classroom activities*) until the emerging themes were generated. Finally, the emerging themes were narrowed down to identify the main themes of the interview responses. Note that we used different colors to group the highlighted statements, as well as to identify which from the following emerging themes were clustered together.

3 Results

3.1 Results of the quantitative analysis

Table 1 shows that Filipino undergraduate students have an overall positive perception of translanguaging (Mean= 3.05, SD= 0.744). This indicates that the majority of the college students who come from different ethnolinguistic groups were in favor of the use of translanguaging in linguistically diverse learning environments. Despite the dominant position of the English language in the Philippines, bilingual and multilingual students from the southern part of the country believe that allowing them to use their full linguistic repertoire at the university and in various social settings is beneficial to them.

Table 1. The perception of Filipino undergraduate students toward translanguaging

	Mean	SD
Perceptions of Translanguaging as a practice	3.06	0.798
Translanguaging should not be avoided by bilinguals.	3.01	.802
Instructors at my university engage in translanguaging.	3.11	.764
Translanguaging is a natural practice for bilinguals.	3.09	.719
Translanguaging indicates linguistic proficiency in your second language.	2.81	.842
Translanguaging is not a disrespectful practice.	3.32	.795
Translanguaging is not confusing for me.	3.02	.863
Perceptions of translanguaging for second language learning	3.01	0.739
Translanguaging help me learn a second language.	3.18	.698
Translanguaging is	2.66	.826

acceptable when you are learning a new language.		
Translanguaging is essential for learning a new language.	3.12	.668
Translanguaging has assisted me in learning a second language.	3.14	.662
Language instructors should not avoid translanguaging because it will prevent second language learning.	2.95	.839
Perceptions of translanguaging in social settings	3.13	0.706
It is okay to engage in translanguaging in social settings.	3.15	.697
I use translanguaging in social settings.	3.08	.730
Translanguaging is socially acceptable.	3.17	.690
Perceptions of translanguaging in higher education	2.99	0.734
It is okay to engage in translanguaging in higher education settings.	2.90	.747
Bilinguals should be able to engage in translanguaging to complete university assignments.	2.75	.730
Translanguaging is acceptable to use within university-level assessments.	2.81	.725
It is appropriate for university instructors to engage in translanguaging.	2.91	.685
Translanguaging by a university instructor is professional.	2.98	.834
I would not feel upset if a university instructor engaged in translanguaging during class.	3.08	.812
If an instructor used translanguaging in class, it would be helpful for the bilingual students.	3.11	.681
Translanguaging helps me engage in conversations with my classmates.	3.17	.696
Translanguaging helps me understand conversations with my classmates.	3.19	.696
Overall	3.05	0.744

Note: 1.00-2.40= negative; 2.50-4.00 = positive N= 1,170

Specifically, Filipino students in Central Mindanao perceive that translanguaging is a good practice (Mean= 3.06, SD= 0.798). Hence, it should not be avoided by bilingual or multilingual students and teachers because they naturally do it. It is a practice in which instructors in the university engage. They also view translanguaging as useful for second language learning (Mean= 3.01, SD= 0.739) because they are permitted to use their mother while developing their skills in the English language. This further means that they think English language learning is possible through translanguaging as opposed to the monolingual approach. The majority of the students agree that it helps them learn their second language and see it as an essential part of learning a new language. Moreover, some believe that even if instructors keep practicing translanguaging, it does not hinder the process of learning the second language.

In terms of the use and function of translanguaging in social settings, Filipino undergraduates positively acknowledge and accept it (M= 3.13, SD= 0.706). Most of them believe that translanguaging is socially acceptable, especially in bilingual or multilingual contexts. The result clearly shows that translanguaging influences students positively when practiced in social settings and interactions.

Lastly, college students in this study have a positive perspective on translanguaging in higher education (M= 2.99, SD= 0.734). They mostly believe that it helps bilingual or multilingual students like them in accomplishing their assignments and in dealing with university-level assessments. Moreover, they consider translanguaging an effective tool for better communication among classmates. The result also implies that they would openly welcome translanguaging when an instructor uses it in-class sessions because they consider it effective assistance to bilingual students.

3.2 Results of the qualitative analysis

Translanguaging is beneficial in classroom activities. One of the reasons why students have positive perceptions of translanguaging is that it is beneficial in classroom activities such as in small group discussions, brainstorming activities, generating examples through translation, recitations, and in-class interactions during a graded discussion like teaching demonstration, as mentioned by the participants of the interview.

> CED#1: I can use my mother tongue every time… I recite, or… sometimes-especially with my course which is… BSEd. I use my mother tongue whenever I... report, when I demo, uhm… every time I am

> asked a question, or I am called… in the class.

> CBA#1: Uhmm, I use my mother tongue sometimes when approaching some members when we have group activities or group reporting.

Translanguaging is an effective tool to communicate and express ideas. Undergraduates perceive translanguaging as an effective tool to communicate wherein some participants responded that the practice helps them in communicating inside the classroom, most especially during discussions and classroom interactions where students may experience difficulty. Similarly, it helps them express ideas in which they can freely share their thoughts without feeling anxious or hesitant. The practice also makes them feel encouraged to share their ideas as one participant expressed.

> CAS#1: For some reason, I think- in uh... some circumstances maybe because... we are having a difficulty when it comes to... Tagalog terms, so- because our teacher can also understand Maguindanaon. So, that is the reason why we are allowed– to use our mother tongue.

> CBA#1: Ahmm, yes, it is a helpful practice in– in a classroom setup. Because it helps us to communicate to other students who are using the same language as us.

Translanguaging allows effective learning. Lastly, translanguaging is perceived to be effective for learning, as it enhances the students' vocabulary, skills, and learning phase in the classroom. Filipino college students also consider translanguaging as a strategy used by teachers to make the learning process easy. In fact, one participant shared that *"it's the strategy of the professor… to help us learn the target language through that"* (CED#1).

> CHS#1: Uhm, I view translanguaging positively… it helps in promoting in enhancing our mother tongue, and it also helps to widen the vocabulary of... ahh all even if others are not speaking... uh... the same language.

> CENCS#1: Ahhh... as from what I can see... ahh... this is beneficial to the students because (...) it really enhances the students. It really hones one's skills, and it gradually helps in learning new languages as well.

In addition, the practice helps them better understand the concepts being taught to them, especially when they translate them mentally into their first language. One participant shared that students who cannot easily follow the discussion can cope with the learning through translanguaging. The participant also added that this practice helps those who have difficulty comprehending the lesson.

> CED#1:(…) the use of mother tongue… can help me as I learn… uhm… the target language which is English. Because, uh… you know… there is… there is… translation in the mind that is happening. And with that, it gives me more understanding and helps me to comprehend.

> CENCS#1: (...) all of us can easily understand if the language that we are using is the one we are all familiar with… I think they have noticed that we can easily understand things easier if we will be using... ahh our own language which is Maguindanaon to another Maguindanaon.

4 Discussion

The result of the quantitative data analysis shows that the majority of Filipino college students positively perceive translanguaging as a practice, for second language learning, in social settings, and in higher education. The overall result indicates a highly positive perception of translanguaging from the majority.

Moreover, according to the participants, translanguaging is a natural practice that occurs in the classroom. Because it usually happens in a learning environment, monolingual educational policies cannot control translanguaging practices (Canagarajah, 2011). It naturally takes place during peer discussions for a deeper understanding of academic tasks and content. This finding can be compared to the study of Fang and Liu (2020) as it also reveals that translanguaging acts as an instruction reinforcement. This approach helps to further clarify any academic instruction through the use of two or more languages. As shared by the participants in the interview, translanguaging is beneficial in classroom activities such as small group discussions, brainstorming activities, generating examples through translation, recitations, and in-class interactions during a graded discussion. It implies that translanguaging encourages students to participate in various classroom activities because they do not feel obliged to speak the target language only. Instead, they are given the opportunity to use their entire linguistic repertoire, setting aside the defined boundaries of the languages that are being utilized and vice-versa (Otheguy et al., 2015).

Most undergraduates view translanguaging as helpful in second language learning and perceive it as a socially acceptable practice. This finding is similar to how American graduate students in the study of Moody et al. (2019) viewed translanguaging in second language learning. They considered it to be beneficial to their second language learning experiences. In this study, the participants explain that the use of translanguaging helps them enhance their vocabulary skills in English. Thus, the present study argues that integrating the practice of translanguaging during the learning process of L2 learning will be a huge help for the students for additional clarity and assistance. Both quantitative and qualitative analyses further support the argument of this research that the monolingual approach has no place in English language teaching and learning. Although there are students who share the traditional view on second language learning, which is speaking English only in an English class, some of them acknowledge the benefits of using their L1 in an L2 learning classroom setting and state that it contributes greatly to their learning process (Tabatabaei, 2019).

Along with that, most students believe translanguaging is an effective tool for understanding conversations with their classmates. This finding, on the other hand, reveals the significance of translanguaging when students are conversing with one another for further precision of exchanged ideas. These perceptions from the quantitative data can be supported by the qualitative data in which students who hold a positive perspective on translanguaging see its effectiveness in communication, classroom discussion, academic work, expression of ideas, and learning as a whole, for they are given the opportunity by their teachers to use two or more languages in the classroom. Similarly, Fang and Liu (2020) found that despite the English-only regulation, the teachers use translanguaging strategies to make content teaching easier, such as concept or language point clarification, comprehension check, and content knowledge localization.

Because students do not have the same level of learning capacity, some students tend to find it hard to catch up with the lessons when there is only one language used. With this, translanguaging becomes an aid for the learners who struggle in class to at least be able to understand and follow the concepts being shared with them. This lessens the possibility of misconception in the class since the opportunity to clarify, elaborate, and translate concepts is fully accessible. In the study of Canagarajah (2011), a student also shared that the utilization of translanguaging helped him to understand clues that are present in a poem as well as the stories being told to them by the teacher.

Moreover, students find translanguaging as an effective tool to express ideas. This reason suggests that translanguaging puts students in a friendly environment wherein they can freely express and contribute their ideas without feeling anxious and hesitant. This shows a high tendency for student involvement because students have the chance to express themselves in any way possible. This supports the findings of Mari and Caroll (2020) where students felt less anxious when they are allowed to incorporate their L1 inside their L2 learning classroom setting. Teachers noted this behavior and decided to create a comfortable learning environment where students are permitted to use Spanish in an English class to better express themselves. Similarly, students' other reason for their positive perception is that it allows them to have better communication with their peers and teachers. This specific reason indicates that translanguaging is also beneficial in social aspects and that it contributes to the relationship built inside the classroom for it reduces possible misunderstandings and barriers caused by the monolingual policy. As a result, the students may feel motivated in learning. As Zhou and Mann (2021) claimed, the integration of a translanguaging approach in both language and content is indeed effective.

Lastly, translanguaging is perceived to be effective in learning. As students shared, this practice helps enhance their skills, broaden their language vocabulary, and is effective in the learning process in general. Therefore, it is safe to conclude that translanguaging benefits the students in learning because it gives them enough access to not just one language, which enables them to participate and take part in the learning process and develop their skills at the same time. Similarly, Zhang (2022) discovered translanguaging as an effective approach to scaffold students to achieve learning growth wherein they become active during their foreign language learning. With these given circumstances, teachers can integrate translanguaging as their teaching approach where they can enhance the performance of their students and establish an effective learning environment inside the classroom.

Taking this into consideration, the use of translanguaging inside the classroom plays a vital role in the learning process of the students as it helps them to fluently communicate their thoughts effectively. In other words, translanguaging helps students by allowing two or more languages to be used during class discussions for them to grasp the lessons and various academic performance instructions accurately. Furthermore, the majority perceived the practice positively because the students see it as an effective tool to use in classroom communication, discussions and brainstorming, academic activities, as well as in building rapport in the learning environment. The results of the quantitative data are supported by the qualitative data, and it clearly shows that the majority of the student participants manifest positive perceptions toward translanguaging in general.

5 Recommendations

Because the majority of Filipino bilingual and multilingual undergraduate students perceive translanguaging positively, this study recommends the integration of translanguaging in higher education. Compartmentalizing the languages of bilingual or multilingual students based on the subjects taken is not viewed by the students as effective. The findings of this study provide the basis of its argument that students at the tertiary level welcome the use of more than one language in the university. Thus, the monolingual approach suggested in Executive Order 210 series 2003 seems inappropriate for a multicultural and multilingual learning environment.

The present study additionally suggests that teachers must be trained to be equipped and familiar with the translanguaging approach. Galante (2020) also saw the need to train teachers so that they will be aware of how translanguaging can be effectively employed in the classroom. It is important to inform teachers that translanguaging is not simple code-switching, but a teaching strategy used to facilitate students' learning experiences. It entails preparation and follows a procedure to ensure that it is appropriately utilized. For example, students are allowed to use their mother tongue or the lingua franca when brainstorming with their classmates but must submit their written output in English. If teachers are well-trained for this approach, students may not be confused or find it difficult to engage in discussions when two or more languages are systematically used.

Yuvayapan (2019) observes that translanguaging practices are hindered because of the expectations coming from the stakeholders, that is, monolingualism is the only approach that should be used in teaching. To prevent these expectations from hindering the implementation of translanguaging, the present study suggests that stakeholders, especially the school administration must be briefed about the importance of translanguaging. Because the school heads have the authority to implement a language policy, they need to be educated about the benefits, purposes, and positive effects of translanguaging in teaching and learning. Doing this will combat misinformation and stigma attached to the use of the local languages inside the classroom setting.

Acknowledging the limitations of the study in terms of the participants, it suggests that future researchers consider surveying the perceptions of translanguaging of other stakeholders in higher education such as college instructors and academic heads in the university. Comparing their perceptions will provide a richer understanding of how translanguaging is viewed in tertiary education. Moreover, further research must be done in various contexts. For instance, it may be noteworthy to replicate this study in a learning environment with monolingual and bilingual students. Monolingual students who share the same learning space as bilingual students may perceive translanguaging differently because they know only one language.

Lastly, other variables may be considered while investigating the perceptions of stakeholders toward translanguaging. Their perspectives may be correlated with language ideologies and language attitudes towards the use of local language and language policy.

References

Abayomi Alase. 2017. The interpretative phenomenological analysis (IPA): A guide to a good qualitative research approach. International Journal of Education and Literacy Studies, 5(2), 9-19.

Adrian Joseph Rivera and Catherine Mazak. 2017. Pedagogical Translanguaging in a Puerto Rican University Classroom: An Exploratory Case Study of Student Beliefs and Practices. Journal of Hispanic Higher Education, (1-15), 153819271773428–.

Angelica Galante. 2020. Pedagogical translanguaging in a multilingual English program in Canada: Student and teacher perspectives of challenges. System, 92(102274).

Anna Mendoza and Jayson Parba. 2018. Thwarted: relinquishing educator beliefs to understand translanguaging from learners' point of view. International Journal of Multilingualism, 1–16.

Ann-Christin Torpsten. 2018. Translanguaging in a Swedish multilingual classroom. Multicultural Perspectives, 20(2), 104-110.

Danping Wang. 2016. Translanguaging in Chinese foreign language classrooms: student and teachers' attitudes and practices. International Journal of Bilingual Education and Bilingualism, 1–12.

Eva Codó PhD. 2008. Interviews and questionnaires. The Blackwell guide to research methods in bilingualism and multilingualism, 158-176.

Eva Mikuska. 2017. The Importance of Piloting or Pre-Testing Semi-Structured Interviews and Narratives. Sage Publications Ltd.

Fan Fang and Yang Liu. 2020. 'Using all English is not always meaningful': Stakeholders' perspectives on the use of and attitudes towards translanguaging at a Chinese university. Lingua, 247(102959).

Fatma Yuvayapan. 2019. Translanguaging in EFL classrooms: Teachers' perceptions and practices. *Journal of Language and Linguistic Studies*, *15*(2), 678-694.

Jorg Blasius and Victor Thiessen. 2001. The use of neutral responses in survey questions: An application of multiple correspondence analysis. Journal of Official Statistics, 17(3), 351-367.

Juliana Zangirolami-Raimundo, Jorge de Oliveira Echeimberg and Claudio Leone. 2018. Research methodology topics: Cross-sectional studies. Journal of Human Growth and Development, 28(3)(356-360).

Khairunnisa Khairunnisa and Iwa Lukmana. 2020. Teachers' attitudes towards translanguaging in Indonesian EFL classrooms. Jurnal Penelitian Pendidikan, 20 (2), 254-266.

Maninder Singh Setia. 2016. Methodology series module 3: Cross-sectional studies. Indian journal of dermatology, 61(3)(261).

Mengqiu Zhang. 2022. A Study of the Attitudes of Chinese Language Teachers towards the Application of Translanguaging during Online Classes in an International School of Bangkok. Language in India, 22(1).

Nataliya V. Ivankova, John W. Creswell, and Sheldon L. Stick. 2006. Using mixed-methods sequential explanatory design: From theory to practice. Field methods, 18(1) (3-20).

Philippine Statistics Authority. 2013. 2010 Census of Population and Housing, Report No. 2A – Demographic and Housing Characteristics (Non-Sample Variables), Cotabato City.

Ricardo Otheguy, Ofelia Garcia, and Wallis Reid. 2015. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. Applied Linguistics Review, 6(3), 281-307.

Rojina Tabatabaei. 2019. Translanguaging in ESL classrooms in Sweden: from the student's point of view. Independent Project.

Rui Yuan and Min Yang. 2020. Towards an understanding of translanguaging in EMI teacher education classrooms. Language Teaching Research, 1362168820964123.

Shannon M. Daniel and Mark B. Pacheco. 2016. Translanguaging practices and perspectives of four multilingual teens. Journal of Adolescent & Adult Literacy, 59(6).

Stephanie Moody, Mahjabin Chowdhury, and Zohreh Rasekh Eslami. 2019. Graduate students' perceptions of translanguaging. English Teaching & Learning, 43(85-103).

Steven Tenny, Grace D. Brannan, Janelle M. Brannan, and Nancy C. Sharts-Hopko. 2017. Qualitative study. StatPearls Publishing, Treasure Island.

Suresh Canagarajah. 2011. Translanguaging in the classroom: Emerging issues for research and pedagogy. Applied linguistics review, 2(1).

Vanessa Mari and Kevin Caroll. 2020. Puerto Rican Teachers' and Students' Beliefs toward Spanish Use in the English Classroom as a Way to Motivate Students. Latin American Journal of Content & Language Integrated Learning, 13(2), 289-311.

Virginia Braun and Victoria Clarke. 2012. Thematic analysis. APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological American Psychological Association, 57–71.

William Alex Edmonds and Thomas D. Kennedy. 2016. An applied guide to research designs: Quantitative, qualitative, and mixed methods. Sage Publications.

Xiaozhou Emily Zhou and Steve Mann. 2021. Translanguaging in a Chinese university CLIL classroom: Teacher strategies and student attitudes. Studies in Second Language Learning and Teaching, 11(2), 265-289.

Zamboanga Chavacano Verbal Aspects: Superstrate and Substrate Influences on Morphosyntactic Behavior

Abee M. Eijansantos
Zamboanga State / Fort Pilar,
College of Marine / Zamboanga
Sciences & Technology / City, Philippines
ameijansantos@zscmst.edu.ph

Jeric B. Ventoza
Ateneo de / Tumaga,
Zamboanga / Zamboanga City,
University / Philippines
ventozajerb@adzu.edu.ph

Rochelle Irene G. Lucas
De La / Taft Ave.
Salle / Manila,
University / Philippines
rochelle.lucas@dlsu.edu.ph

Ericson O. Alieto
Western / Baliwasan,
Mindanao / Zamboanga City,
State University / Philippines
ericsonalieto@gmail.com

Abstract

It was hypothesized that Zamboanga Chavacano verbs exhibit variation in the need for a verbalizer prior to verbs whose origin is not the superstrate Spanish. To address the hypothesis, 200 verbs from the 500,000-word Zamboanga Chavacano corpora were analyzed using AntConc 3.5.8 (Windows) 2019. Additionally, verbs from different origins were subjected to analysis. Taken from the Composite Dictionary of Riego de Dios (1989), 100 verbs which were Spanish originated and another 100 that were non-Spanish were subjected to elicitation among 104 native speakers of Zamboanga Chavacano. The data revealed that the same case was observed for Zamboanga Chavacano as the Cotabato variety, but the verbalizer had the tendency to cliticize with the perfective and imperfective aspects, while only deletion took place for the contemplative aspect. Additionally, verbs that are inflected required the verbalizer no regardless of their origins.

Keywords: Zamboanga Chavacano, Spanish verbs, aspect marker, verbalizer

1. Introduction

The Chavacano language is a Creole language (Lipski, 1987, 2012; Barrios, 2006; Wolff, 2006; Meyerhoff, 2008; Steinkrüger, 2008, 2013; Paz, Hernandez, & Peneyra, 2010; and Porras, 2013) that is Spanish-based (Holm, 2001). Creole formation occurs where both the superstrate and substrate languages contribute to the creole's structure (Crowley, 1997; Holm, 1988). Further, creoles emanate from the source languages which influence their grammar, but the features are adopted in a rather "less complex rendition" (McWhorter, 2018, p.18).

Premised on the preceding claims, Chavacano as a Creole is expected to exhibit traces from its source languages: Spanish (Lipski, 1986 as cited in Barrios, 2006), as its superstrate language, and some Philippine-type languages such as Filipino and Cebuano (Barrios, 2006) as its substrate languages. As such, Chavacano is a Creole language that exudes syntactic influences from its source languages. One of these observations is that of Riego de Dios & Otanes (1989) of the Chavacano variant spoken in Cotabato (henceforth Ct) which is said to have been greatly influenced by the variety spoken in Zamboanga. They claimed that the Ct's verbs behave varyingly which is dependent on the language source of the verb. Verbs whose source is a Philippine language or English, a verbalizing marker—*man*—is necessary to be prefixed to a verb stem prior to the prepositioning of the aspect marker—*ya* for past/punctual, *ta* for the present/durative aspect or *ø/ay* for future—while no prefixing is syntactically needed when the verb stem is Spanish-derived.

Thereby, it was hypothesized that (1) Zamboanga Chavacano (hereafter ZCh) exhibits an identical syntactic behavior in the selection of the aspectual markers where a verbalizing marker is necessary for non-Spanish-derived verbs to permit the aspectual marker in the sentence, while it is unnecessary for Spanish-derived ones. Also, premised on the primary observation, (2) the verbalizer is suspected to make the aspect marker the host for it to cliticize with.

To address these hypotheses, (1) 104 native speakers of ZCh were requested to be the participants of the syntactic construction being explored, and (2) a corpus analysis of 200 ZCh verbs from the collected corpora ranging from radio scripts, bible translations, newspapers and others was likewise analyzed via AntConc 3.5.8 (Windows) 2019.

Research Questions

To help address the hypotheses above, the following research questions served as guides in the entire conduct of the research:

1. In what morphosyntactic environment is *man* required in a ZCh sentence?
2. What morphosyntactic phenomenon is taking place in the difference in the following constructions:
   yan and *ya man,*
   tan and *ta man,*
   man and *ay man*?

2. Review of the Related Literature

2.1 ZCh Creole

The Chavacano language is creolized (Barrios 2006; Holm, 2001; Lipski 1987, 2012; Meyerhoff 2008; Paz, Hernandez, & Peneyra 2010; Porras 2013; Steinkrüger 2008, 2013; Wolff 2006). Barrios (2006) explained that Zamboanga Chavacano is one of the Philippine Creole Spanish (PCS) varieties, while Lipski (1987) explicated that it is the only surviving variant. Holm (2001) argued that among the PCS, ZCh variant is the creole that is likely to flourish, and in fact, has the largest population with roundabout half a million speakers (Fernández, 2001; Grant, 2007; Lipski, 2001).

2.2 Tense and Aspect

Tense is a grammatical element intended to situate events, states, or actions in relation to time, while aspect concerns with the inner structure of an action taking place *'at a given time'* (Celce-Murcia & Larsen-Freeman, 2015, p. 106). Similarly, aspect is referred to as a grammatical feature that is referred to the time dimension, but is not associated with a particular point in time which is a property of tense (Payne, 2011). Similarly, Comrie (1976, 1985 as cited in Viti, 2016) explained that tense typically pertains to the time of an event as 'present', 'past', or 'future' in the context of the time of speaking. Aspect, however, pertains to a situation in itself, which can be regarded as 'perfective', imperfective', or 'resultative'. Further, tense is subsumed under a deictic category in that it locates situations in relation to time (Summer Institute of Technology (SIL), 2021; Hermont & Martin, 2020), while aspect related to the situation's internal time (Hermont & Martin, 2021). Also, tense which relates to time is contrasted with aspect that is analyzed as a grammatical element related with an action's internal structure occurring at any time (Murcia & Freeman, 2008). More particularly, the time articulated in the English verb is based on a certain *point in time*, and premised on *flow of time* in Filipino (Schachter & Otanes 1972, as cited in Ceña & Nolasco, 2011, p.75).

2.3 Creole Verbal Morphology

The verb phrase of creoles has been the focus of creolists, and has been instrumental in distinguishing creoles from noncreoles (Holm 2004). Moreover, TMA (tense, aspect, mood) has been a topic that has received much attention in creole studies (Frank, 2004). In fact, Velupillai (2015) claimed that creoles constitute TMA markers with a definite pattern. It is assumed that creoles will have one marker each for tense, aspect, and mood. Additionally, Holm (1988) argued that these language types contain little to no inflectional morphology. Contrarily, Plag (2002) argued that the preceding notion relative to creole morphology has to be foregone or prodigiously revised as creoles carry morphological problems as the non-creole languages do.

Though not exclusively isolating in structure, pidgins and creoles considerably form part of the world's isolating languages. Moreover, pidgins and creoles differ from their source languages in their trademark simplicity in morphology. Truly, multilingual contact in a faulty learning environment offshoots an isolating morphology. With caution, the author admitted that the foregoing morphological simplicity should not restrict the creolist in analyzing a complexity when faced with one in a creole. In fact, cliticization is even a feature that can possibly take place in these types of languages (Crowley, 2008).

Winford (2006) analyzed the TMA markers of some creoles as isolating. Similarly, Bakker (2002) suggested that Tok Pisin has isolating TMA markers, and regarded Sranan to have the same. There are those that appear to exhibit an inflected morphology in the TMA markers like the analysis of Kouwenberg (1994) of the Berbice Dutch Creole. The perfective and imperfective aspects are encoded as morphologically inflectional. Simply put, the morphology of some creoles permits inflection or affixation, meaning, creoles are not solely morphologically isolating. As such, Farquharson (2007) analyzed the Jamaican Creole as having a progressive aspect affix since it is obligatory when eventive verbs are used. These elements cannot take independent stress and separate with the verb using other elements. Likewise, the progressive marker, when separated from the verb, is ungrammatical.

2.4 ZCh Aspect Markers

The ZCh morphosyntax is in dire need of a linguistic description. Some authors like Riego de Dios and Otanes (1989) described a variety of Chavacano that is related to ZCh, the Ct. In their analysis of the aspect markers of Ct, *ya***,** *ta* and *ay* are analyzed as tense unbound morphemes for past, present/durative, and past/punctual respectively. These markers according to them are used without the *–r* for Spanish infinitives. For Philippine source verbs, the verbalizer *man* is prefixed before a verb to permit the aspect marker to figure in the utterance. Similarly, Lipski and Santoro (2007) argued that the markers *ta, ya*, and *ay* are used to refer to different meanings and are free-standing morphemes.

Barrios and Bernardo (2012) presented the perfective aspect marker, as in *ya ase apas* 'chased' and *ya kay* 'fell', with the unbound aspect marker *ya*. Likewise, Barrios (2006) claimed that the tense markers used in ZCh are *ya* for punctual and perfective; *ta* for habitual, durative, present or past imperfective; and *ay, ey* or *el* for future. Likewise, these markers are free morphemes, for instance, *Ya come el perro conel pescao* 'The dog ate the fish.' The same can be observed with Santos (2010) and Camins (1999), while Porras (2013) presented sentences suggesting the isolating feature of the Chavacano verbs.

Steinkrüger (2008) claimed that ZCh has the following TMA markers: *ta* for imperfective, *ya* for perfect(ive), *ay* for irrealis, and *kaba* for completive. In his analysis, though, differing from the preceding authors, the preverbal markers behave as prefixes morphologically. This analysis is anchored on the similarity it exhibits with many Philippine languages.

As far as the syntactic element *man* is concerned, only Riego de Dios and Otanes (1989) and Lipski and Santoro (2007) investigated this element, calling it a verbalizer or a derivational suffix respectively. For the purpose of this paper, the term *verbalizer* is favored.

3. Research Methodology

3.1 Participants/ Informants

This study adopted the definition of a native speaker from the linguistics viewpoint. Specifically, in a sociolinguistics approach, Davies (1991) defined a native speaker in terms of social identity using the criteria such as the early acquisition and the continuous cultural/oral tradition. In the lens of modern linguistics, Chomsky (2014) explained that a native speaker possesses the authority over the system and structure of their native language. Furthermore, Lee (2005) itemized the six criteria to best identify a native speaker: (1) the acquisition of the language in early childhood; (2) the intuitive knowledge on the language they use; (3) the ability to speak the language with fluency and spontaneity; (4) the capacity to communicate the language with competence and the flexibility to utilize the language in various set-ups; (5) the recognition of the community of the speakers' nativeness; and (6) the absence of a foreign accent.

As such, the researchers adopted the inclusion criteria anchored from Eijansantos et al. (2021): (1) they had to have been using ZCh as a medium of communication from childhood until the present; (2) they had to recognize themselves as a native speaker of the language who had acquired the language at the age of or earlier than seven; (3) they had to be at least 18 years old; and (4) they had to be at least in the Senior High School.

Thereafter, the researcher-made-Chavacano test was taken by the 104 participants who passed the requisites on the aforesaid criteria. The participants were required to gain at least 70%

or 10 out of 15 as their result to ensure that they are native speakers of ZCh. Moreover, purposive and convenience sampling techniques were used.

3.2 Data Collection Procedure

The data were collected in two ways. The first of the data collection procedures was the corpus analysis. In this investigation, the researchers utilized the ZCh corpora, a research material that contained 500,000 Chavacano words whose primary goal was to accumulate ZCh verbs which became the subject of the corpus analysis in determining the morphosyntax of the varying ZCh verbal aspects. Subsequently, the ZCh corpora were subjected to Antconc 3.8.5 (Windows) 2019 to facilitate the identification of the verbs premised on the hypothesis formulated that the source language influences a grammatical variation in the ZCh verbal string. Through the aforementioned software, the researchers were able to identify the verbs that use *yan* with 776 hits and the verbs that use the aspect marker *ya* with 23, 887 hits. With the huge number of hits from both markers, the decision on the reduction of the number to an attainable margin was put forth. Thus, only 100 verbs were selected from each generated result, totaling 200 verbs which were subjected to analysis. To corroborate the lexical origin of the verbs–whether these verbs were Spanish or non-Spanish in origin—five different dictionaries were utilized: (1) Composite Dictionaries developed by the and the Linguistics Society of the Philippines (LSP); (2) UP Diksyonaryong Filipino; (3) Online Cebuano Dictionary; (4) Online Spanish Dictionary; and (5) Cambridge Spanish Dictionary. All these dictionaries contributed to the identification of the verbs' origins that aided the corpus analysis.

The inclusion criteria were intended to filter the informants for the elicitation task which was the second data collection procedure. The researchers ascertained that participants were selected using the abovementioned criteria before they were subjected to the elicitation task, and the test was consentaneously administered to them. With their consent, the instrument which was generated via a Google form that was individually sent to them. A part and parcel of the inclusion criteria was the researcher-made-Chavacano test. In the said test, the participants were required to achieve at least 70% or an equivalent score of 10 out of 15 as the total score.

Elicitation is a method to obtain data from speakers of a language that may either be actual utterances or judgments about the acceptability of an utterance (Crystal, 2011). The elicitation task in this study was a sentence completion type of test that required the participants to use the correct aspect marker *ya* or *yan* in the sentence. This task sought to clarify the difference among the verbs and intended to show where this difference lies pertinent to the aspectual markers. This test contained 40-item ZCh sentences examined by the 104 participants, who were identified and who qualified as bonafide ZCh native speakers. The verbs that were used for the elicitation task were taken from the Composite Dictionary of Riego de Dios (1989) in that it clearly showed and differentiated verbs that were either ZCh or Ct. If a verb was uniquely ZCh or Ct, the dictionary indicated so. Furthermore, as this paper was premised on the hypothesis that Spanish verbs require a separate element or morpheme in their configuration, verbs that were Spanish in origin and those that were otherwise not were selected from the dictionary as it clearly specified a verb's source language. Thus, the participants' judgement on the sentences as ZCh native speakers would serve as empirical evidence to justify the syntactic environment the verbalizer *man* is required in ZCh and to elucidate the morphosyntactic phenomenon that takes place in the distinction between the construction of *yan* and *ya man*.

Moreover, the data collected from the corpus were comparatively more recent than the lexical items taken from the Composite Dictionary—that is, the corpus has quite recently been put together, while the latter was published in the late 1980s. Although there may be some discrepancies as far as time periods were concerned, the dissimilarity in time periods further strengthens the claims made in this paper in that consistency in the morphosyntactic behavior of the verbs has been succinctly determined from the two data gathering procedures. The verbal aspect markers behave contrarily when the verb is an uninflected Spanish verb. Moreover, the analysis was carried out two-way—corpus analysis and elicitation—for triangulation or corroboration purposes where the data in one data gathering procedure reinforce the ones in the other, and vice versa.

3.3 Data Analysis

The corpus analysis was carried out via the investigation of the source language of each of the verbs in the hits generated via AntConc. Apart from this, the verbs in the hits analyzed were put under scrutiny relative to their morphological constituents whereby the affixation—wherever applicable—was teased out. Also, the grammatical category of each of the verbs in ZCh sentences in their original form from where they emanate was likewise considered. Each of the hits taken as entries in the data analysis was laid bare to be analyzed pertinent to the need to use either *ya* or *yan* in a construction. In summary, each of the hits for both the verbs with *ya* and those with *yan* were dissected linguistically.

To execute the analyses of the data gathered from the participants via the elicitation task, a statistical treatment that was done through the expertise of a statistician was carried out. In the treatment, a binomial test at 50% proportion with a significance level of .05 was employed. This certainly helped foster the significant difference in the participants' elicitation task responses on their choice of the proper use of the aspect markers *ya* or *yan* in the sentences. The significant difference, grounded in the results in the statistical treatment, evidently affirmed the morphosyntactic phenomenon of the above-mentioned aspect markers in their syntactic environment thereby validating the researchers' hypothesis. Further, analogous to the analysis executed for the corpus data, each of the sentences was teased apart in the lens of the source language of the verbs. Additionally, other potentially relevant linguistic features were scrutinized, for instance, the lexical item's verbal morphological algorithm and their grammatical category in the language of origin. Simply put, the constructions were subjected to an in-depth linguistic exploration.

As far as the apparent fusion that takes place between the aspect marker *ya* and the verbalizer *man*, the teasing out was done via the tree diagrams patterned from Nolasco and Ceña (2011) of their description of the Tagalog syntax. Following this line of thinking aided the analysis of the other two aspects in ZCh sentential structure in that in principle, highly identical processes apply for them—that is, in the imperfective and contemplative aspects. This was carried out with cautious and mindful examination for potential variation that may be either apparent or otherwise concealed.

4. Results and Discussion

4.1 Corpus

Of the first set of the 100 selected verbs from the corpora, each one is a Spanish verb that potentially means that they are Spanish in origin, too. All these verbs use the *ya* aspect marker without the verbalizer *man* in the construction. The following examples seek to clarify the verbal aspect construction among verbs that are Spanish in origin:

(1) *Ya* *agarra*
PRF 'grasp or take hold of'
'grasped or took hold of'
(2) **Ya* *man* *agarra*
PRF VBLZ 'grasp or take hold of'
'grasped or took hold of'
(3) **Yan* *agarra*
PRF=VBLZ 'grasp or take hold of'
'grasped or took hold of'

Notice that when the verbs above are succeeded with the verbalizer *man* or its cliticized form (see items 1 and 2), an ungrammatical construction yields. These verbs, when compared with the ones below, are succinctly different in the necessity of the verbalizer *man* which surfaces as a clitic attached with the *ya* aspect marker. A clitic is a morpheme, that may have a nonclitic alternant, has syntactic characteristics exhibited by a word, but manifests a behavior of being phonologically attached to another word. Specifically, an enclitic is a clitic that joins phonologically at the end of a preceding lexical item to form one unit (SIL, 2022).

The verbs in the examples below (see items 4 and 5) which were taken from the other set of 100 verb strings from the corpora contained the aspect marker *yan/ya (man)* + verb. These verbs are analyzed as requiring the verbalizer *man* in the verbal string which is analyzed as a construction that is an offshoot of cliticization (i.e. Ya + man → yan). This resonates with Crowley's (2008) argument that this process is permissible in creole languages and parallels what Kouwenberg (1994) has claimed about Berbice Dutch Creole which exhibits a morphological inflection in its verbal structure. It further fortifies what Farquharson (2007) has remarked about Jamaican Creole's utilization of an aspectual affixation. All these findings substantiate McWhorter's (2005) assertion that creoles' complexity from an original state later on flourishes. This process of the verbalizer cliticizing with the aspect marker contrasts to some degree with Winford's (2006) delineation of TMA markers among creoles as isolating which are exemplified in the data he presented and Bakker's (2002) presentation of Tok Pisin's isolating TMA algorithm. Due to the fact that cliticization remains not obligatory up until now (i.e. *ya man, ta man),* Winford's (2006) delineation and Bakker's (2002) presentation are likewise validated. Further, the majority of the verbs in the other set of 100 select ones are non-Spanish in origin, for example, *Yan apas* 'to follow/ catch up with' from Ceb, *Yan decide* from Eng, *Yan bulabog* 'to wreak havoc/to disrupt peace' from Tag. The same case holds true for the verbs that are Spanish in origin but are inflected in the ZCh sentence, or are not verbs but are either nouns or adjectives in that language of origin. The following handpicked examples are provided to clearly present the need of a verbalizer in the verbal construction:

(4) *Yan* *apas*
PRF= VBLZ 'follow/catch up with'
'caught up with'
(5) *Ya* *man* *apas*
PRF VBLZ 'follow/catch up with'
'caught up with'
(6) **Ya* *apas*
PRF 'follow/catch up with'
'caught up with'

Constructions with an aspect marker and the clitic verbalizer (see item 4) like the one where the verbalizer is a free morpheme (see item 5) are all grammatical constructions. In contrast, those without the verbalizer yield an ungrammatical string as in item 6. It should be noted that *apas* is Ceb. Identically, the same morphosyntactic behavior is exhibited by the Tag verb *bulabog* 'wreak havoc,' Eng verb *explain*, Spa noun *encuentro* 'meet', and the Spa verb *hila* 'pull' but is inflected in its ZCh structure *hilaan* 'pull something from each other'. All these are consistent with the hypothesis that verbs which are Spanish in origin do not require the verbalizer *man,* but those that are otherwise non-Spanish do so. A novel analysis has to be clarified, though. Verbs in ZCh sentence which are Spanish in origin, but are not actual Spanish verbs or are lexical items belonging to other grammatical categories, require the verbalizer. Similarly, inflected Spanish verbs in ZCh sentence need the verbalizer to render the sentence grammatical.

In other words, comparing the verbs with the *ya* aspect marker and the ones with the *yan* aspect marker, only verbs that originate from Spanish and are actual Spanish verbs are the ones that do not necessitate the verbalizer *man* prior to the ZCh verb string. This foregoing finding sheds light to the first research question: *In what syntactic environment is* ***man*** *required in a Zamboanga Chavacano sentence?*

4.2 Elicitation Task

All the statistical decisions are significant as a result of the binomial test as far as the elicitation task is concerned. In other words, the significant difference in each of the items is indicative of the choice of the majority that is regarded as the preferred grammatical construction. In each of the sentences in the elicitation task, the respondents had to choose between *ya* and *yan*. The following are selected items taken from the instrument used for the elicitation task.

(7) ____ *abla le* *el* *deberasan.*
tell 2sg.NOM DET truth
's/he told the truth'
(8) _____ *agarahan* *sila* *mano.*
hold-RECP 3pl.NOM hand
'They held each other's hands.'
(9) _____ *saguan sila* *na* *baroto.*
oar 3pl.NOM LOC boat
'They rowed in the boat.'

(10) _____ *bitay el chonggo na pono*.
Hang DET monkey LOC tree
'The monkey hanged by the tree.'

From the 11 items in the instrument, whose verbs are Spanish in origin, the *ya* aspect marker gains the upper hand (see example item 7 above). Notice, however, that for sentence number 8 above taken from the 10 items in the instrument, the verbs, although Spanish in origin, require the *yan* aspectual marker. Apparently, when the Spanish verbs figure as an inflected construction in ZCh, *yan* is preferred. A similar behavior can be observed among Spanish lexical items that are not verbal—that is, adjectives and nouns (i.e. sentence 9), which, albeit Spanish in origin, detach from the necessity to co-occur with *ya* as *saguan* 'oar' is a noun, and as an extension, it is believed that an identical case holds true for other lexical items belonging to other syntactic categories. Moreover, the verbs whose origin is non-Spanish oblige the *yan* aspectual marker in a ZCh sentence as can be seen in sentence 10 above which contains a verb that comes from Hiligaynon. The same case has been found from the verbs that emanate from English, Tagalog, Cebuano, and those that are local to ZCh. All these findings are analogous to Riego de Dios' (1989) description of Ct that requires a verbalizer to morphosyntactically co-occur with the verb prior to forming the whole verb phrase string; however, the findings extend the description in terms of the analysis of the verbalizer as a free morpheme and can potentially transition into a clitic.

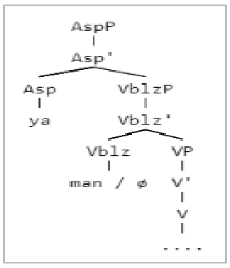
The findings in this component of the investigation corroborate those in the corpus analysis where only uninflected Spanish verbs use the *ya* aspect marker in ZCh sentences. Hence, verbs that are non-Spanish in origin and those that are inflected in ZCh or are Spanish in origin but are not Spanish verbs require *yan*. As in the comparison of the related literature with the findings in the corpus analysis section, the same can be remarked pertinent to the authors and their respective works mentioned.

Furthermore, the findings in this section address the first research question which is *In what syntactic environment is* ***man*** *required in a Zamboanga Chavacano sentence?*

At an initial glance, the supposed counterexamples are thought to pose an issue in the analysis that can dismantle the integrity of the analysis in this paper: *Yan desgrasya* 'met an accident,' and *Yan disciplina* 'disciplined,' have a penultimate stress suggesting that they are Spanish nouns. When the stress falls on the final syllable, they are verbs. Further, *Ya taya* 'bet/cut, chopped' poses another conundrum in the analysis in that the verb originates from Tagalog. However, Spanish has influenced not only ZCh but potentially other Philippine languages as well. Furthermore, the verb *taja* 'to slice' is a Spanish verb, and a verb in Asturian *taya* 'to pay a quantity of money under feudalism' or 'to cut' exists. Asturian is said to have been in contact with Spanish since the 14th century (Barnes, 2013), and that language contact brings about language mixture, change, or the worst case scenario, death (Thomason, 2001).

Now, to address the *second research question*, the following analyses are in order as premised on the syntactic analysis of Nolasco and Ceña (2011) of Tagalog.

Figure 1.



For the ZCh verbs that are non-Spanish in origin, are Spanish but are inflected in the ZCh sentence, or are Spanish but rather belong to a different syntactic category apart from verbs, the verbalizer head rises to the aspect head in order to permit encliticization of the verbalizer with the aspect marker ensuing in the cliticized *yan*. When it opts not to cliticize, the surface form figures as *ya man*, and thus no movement takes place. A similar pattern can be noticed in the imperfective aspect *ta* where the head verbalizer can move up to cliticize with the aspect marker. Depending upon the preference to cliticize or not, the surface structure can either be *tan* when cliticized or *ta man* when otherwise. The preceding validates Crowley's (2008) remark relative to cliticization, Riego de Dios' (1989) description of Ct variety necessitating a verbalizer for a certain type of verb, and McWhorter's (2005) perspective that creoles manifest complexity in time. The fact that creoles are not exclusively isolating in its grammatical architecture is validated, paralleling Kouwenberg (1994) and Farquharson (2007) of their remark in regards Berbice Dutch Creole's inflectional proclivity and the Jamaican Creole's affixational feature, only ZCh exhibits cliticization and deletion—as tackled succeedingly. Further, a rather varying observation is seen when the verb is in the contemplative aspect *ay*. When the head verbalizer raises to the aspect head position, the aspect marker a*y* is deleted instead of the verbalizer cliticizing with it. This may be ascribable to phonological grounds—that is, the shortened versions are parallel in their syllable structure and in their form in general: *yan* for perfective; *tan* for imperfective; and simply ~~*ay*~~ *man* for contemplative.

As in the observation above, for the verbs that are Spanish in origin and are actual Spanish verbs in ZCh, the head verbalizer also has the tendency to rise to a higher position in the structure; however, no cliticization is permissible. In this configuration, since the verbalizer is tacit, no syntactic changes are evident in the surface structure, thereby no cliticization is possible despite the movement of one element to a higher position. Cliticization with the tacit verbalizer is not permissible in both the imperfective and perfective aspects—*ya* and *ta*—due to the phonologically unpronounced verbalizing allomorph ø. Be that as it may, there is evidence that movement indeed is a system occurring in the verbal string. This is apparent in the contemplative aspect *ay* where deletion is possible. It is grammatical to say (*Ay) kome le maniana* 'She will eat tomorrow,' where the sentence in the contemplative aspect can either be with the aspect marker or otherwise not. Hence, when an element moves up, deletion of the contemplative aspect marker occurs.

To succinctly present the morphosyntactic behavior of the verbal string in ZCh, the following is provided to simplify the preceding discussions:

ya + man → *ya man* or its cliticized form *yan*
ta + man → *ta man* or its cliticized form *tan*
ay + man → *ay man* or its deleted form *man*

5. Conclusion

The verbalizer *man* is unessential when the ZCh verb is an uninflected Spanish verb in the ZCh utterance. The verbalizer *man* is necessitated in the verb string in the following: with non-Spanish verbs; with Spanish verbs that are inflected in the ZCh sentence; or with a Spanish noun, adjective or—extraneously—with any nonverbal syntactic category used as verbs. Moreover, the aforesaid verbalizer can cliticize with the perfective and imperfective aspects and thus can appear in the surface form as a cliticized and shortened form. In the contemplative aspect, however, the case is noticeably at variance in that instead of cliticization, the aspect marker undergoes deletion which is attributable to phonological motives. Additionally, the hypotheses established at the outset of the paper have been confirmed: ZCh behaves identically with Ct, and novel findings also have been brought to light, for instance, the nonexclusivity of *man* to non-Spanish lexical items and its cliticizing behavior.

In the purview of education, possessing an understanding of the tacit structural complexities of the grammar of one's native tongue is an advantageous starting point to transition into learning the grammatical system of any second language—in the case of the Philippine educational system—English and/or Filipino. This is grounded on the mother tongue-based multilingual education, a concern that remains unheeded as far as the policymakers are concerned despite the voluminous research tremendously backing up MTB-MLE's beneficial contribution to the success of the learners of L2.

In addition, a study that involves language is concomitantly linked with culture; premised on the foregoing, these current findings have a bearing in culture studies as there is an inherent intertwining between them. Hence, it is assumed that an investigation of the ZCh language is likewise a scrutiny of its speakers' rich culture.

Abbreviations

1 → first person	pl → plural
2 → second person	PRF → perfective
3 → third person	sg → singular
Ceb → Cebuano	Tag → Tagalog
Eng → English	VBLZ → verbalizer
Loc → Location	VP → Verb Phrase

Bibliography

Almario, V. S. (2010). UP diksiyonaryong Filipino: Binagong edisyon. *Quezon City: UP sentro ng wikang Filipino*.

Bakker, P. (2002). Pidgin inflectional morphology and its implication for creole morphology. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of Morphology*, 3-34. New York, Boston, Dordecht, London, & Moscow: Kluwer Academic Publishers

Bakker, P. (2008). Pidgins versus creoles and pidgincreoles. In: Kouwenberg, S., Singler, J.V. (Eds.), *Handbook of Pidgin and Creole Studies* (pp. 130-157). Blackwell

Bakker, P., Daval-Markussen, A., Parkvall, M., Plag, I. (2011). *Creoles are typologically distinct from non-creoles. Journal of Pidgin and Creole Languages,* 26, 5–42.

Barnes, S. (2015). Perceptual salience and social categorization of contact features in Asturian Spanish. *Studies in Hispanic and Lusophone Linguistics*, *8*(2), 213-241.

Barrios, A. L. (2006). Austronesian elements in Philippine Creole Spanish. *Philippine Linguistics Journal, 37*, 34-49.

Barrios, A. L., & Bernardo, A. B. I. (2012). *The acquisition of case marking by L1 Chabacano and L1 Cebuano learners of L2 Filipino: Influence of actancy structure on transfer. Language and Linguistics, 13.3*, 499-*521.*

Camins, B. S. (1999). *Chabacano de Zamboanga Handbook and Chabacano-English-Spanish Dictionary.* Zamboanga: Claretian Publ.

Cebuano Dictionary. (2022). Pinoydictionary.com. https://cebuano.pinoydictionary.com/

Celce-Murcia, M., & Larsen-Freeman, D. (2015). *The grammar book: An ESL/EFL teacher's course*. Boston, MA: Heinle & Heinle.

Ceña, R. M., & Nolasco, R. M. D. (2011). *Gramatikang Filipino: Balangkasan.*Quezon City: The University of the Philippines Press.

Payne, S. G. (2011). *The Franco Regime, 1936–1975*. University of Wisconsin Pres.

Crowley, T. (1997). *An introduction to historical linguistics*. Auckland, Oxford, and New York: Oxford University Press.

Crowley, T. (2008). Pidgin and Creole Morphology. In Kouwenberg, Silvia & Singler, John Victor (eds.), *The Handbook of Pidgin and Creole Studies,* 74-97. Singapore: Blackwell Publishing Ltd.

Crystal, D. (2011). *A dictionary of linguistics and phonetics*. John Wiley & Sons.

Davies, A. (2003). The native speaker: Myth and reality. https://shorturl.ae/wP2Hg

DeGraff , M. (2003). Against creole exceptionalism. *Language, 79*(2), 391–410.

Eijansantos, A. M., Alieto, E. O., Emmanuel, M. S., Pasoc, M. G. O., & Bangayan-Manera, A. (2021). Interspeaker variation in the negated perfective aspect of Zamboanga Chavacano. Linguistics and Culture Review, 5(S3), 287-309. https://doi.org/10.37028/lingcure.v5nS3.1528

Farquharson, J. T. (2007). Typology and grammar: Creole morphology revisited. In Ansaldo, Umberto & Matthews, Stephen & Lim, Lisa (eds.), *Deconstructing creoles* 21-38. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Fernández R. M. (2001). ¿Por qué el Chabacano?. *Estudios de sociolingüística: Linguas, sociedades e culturas, 2*(2), 1-12.

Frank, D. B. (2004). Creoles, contact, and language change. In Escure, Geneviève & Schwegler, Armin (eds.), TMA and the St. Lucian Creole verb phrase, 237-258. Amsterdam & Philadelphia: John Bejamins Publishing Company.

Grant, A.P. (2007). *Some aspects of NPs in Mindanao Chabacano: Structural and historical considerations.* In M. Baptista & J. Gueron (eds.), *Noun Phrases in Creole Languages: A multi-faceted approach.* Amterdam: Benjamins.

Hermont, A. B., & Martins, A. L., (2020). Tense, aspect, mood and modality. *SCRIPTA* 24(51), 27-45.

Holm, J. (1988). *Pidgins and Creoles.* New York and Melbourne: Cambrigde University Press.

Holm, J. (2001). *Chabacano versus related creoles: (socio-) linguistic affinities and differences. Estudios de Sociolinguistica, 2*(2),69-93.

Holm, J. (2004). *An introduction to pidgins and creoles.* Cambridge: Cambridge University Press.

Holm, J. (2006). *Portuguese- and Spanish-based creoles and typologies*. PAPIA 16: 53-61.

Kouwenberg, S. (1994). A grammar of Berbice Dutch creole. [Mouton Grammar Library 12.] Berlin: Mouton de Gruyter.

Kouwenberg, S., & Singler, J. V. (2008). *The handbook of pidgin and creole studies.* Chichester, England: Blackwell Publishing Ltd

Lipski, J. M., & Santoro, S. (2007). Zamboangeño Creole Spanish. In John Holm, & Peter Patrick (Eds.), *Comparative creole syntax. Parallel outlines of 18 creole grammars* (pp. 372-398). Plymouth: Battlebridge Publications.

Lipski, J. M. (1987). *Chabacano/Spanish and the Philippine linguistic identity*. Unpublished manuscript.

Lipski, J. M. (2001). *The place of Chabacano in the Philippine linguistic profile. Estudios de Sociolingüística* 2(2) 2. 119-163.

Lipski, J. M. (2012). *Remixing a mixed language: The emergenc of a new pronominal system in Chabacano (Philippine Creole Spanish). International Journal of Bilingualism, 17*(4), 448-478.

Lee, J. J. (2005). The native speaker: An achievable model. *Asian EFL Journal*, *7*(2), 152-163.

Leong, S.K., Tsung, L.T.H., Tse, S.K., Shum, M.S.K., & Ki, W.W. (2012). *Grammaticality judgment of Chinese and English sentences by native speakers of alphasyllabary: a reaction time study.* International Journal of Bilingualism, 16(4), 428-445.

McWhorter, J. (2001). *The world's simplest grammars are creole grammars. Linguistic Typology* 5(2–3): 125–166.

McWhorter, J. H. (2005). *Defining Creole.* Oxford: Oxford University Press.

McWhorter, J. H. (2018). *The creole debate*. Cambridge University Press.

Meyerhoff, M. (2008). *The Handbook of Pidgin and Creole Studies.* In S. Kouwenberg & J. Singler (eds.), West Sussex: Blackwell Publishing Ltd.

Paz, C. J., Hernandez, V. V., & Peneyra, I. U. (2010). *Ang Pag- aaral ng Wika.* Quezon: The University of the Philippines Press.

Plag, I. (2002). *Introduction: The morphology of creole languages.* In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of Mophology,* 1-2. New York, Boston, Dordecht, London, & Moscow: Kluwer Academic Publishers.

Payne, S. G. (2011). *The Franco Regime, 1936–1975*. University of Wisconsin Pres.

Porras, J. E. (2013). *Noun phrase marking in Chabacano (Philippine Creole Spanish): A comparative perspective. California Linguistic* Notes, 38(1), 122-142.

Riego de Dios, M.I.O., & Otanes, F.T. (1989). *Studies in Philippine linguistics.* Linguistics Society of the Philippines and Summer Institute of Linguistics.

Santos, R. A. (2010). *Chavacano de Zamboanga: Compendio y diccionario: Chavacano-English, English-Chavacano*. Zamboanga: Ateneo de Zamboanga University Press.

Steinkrüger, P. (2006). *The puzzling case of Chabacano: Creolization, substrate, mixing and secondary contact.* Presented at tenth international conference on Austronesian linguistics, Puerto Princesa City, Palawan, Philippines. Retrieved from http://www.sil.org/asia/philippines/ical/papers.html

Steinkrüger, P. O. (2008). *The puzzling case of Chabacano: Creolization, substrate, mixing and secondary contact. Studies in Philippine Languages and Cultures, 19,* 142-157.

Steinkrüger, P. O. (2013). Zamboanga Chabacano structure data set. In S. M. Michaelis, P. Maurer, M. Haspelmath, & M. Huber (Eds.), *Atlas of Pidgins and Creole Languages Structures Online.* Leipzig: Max Plank Institute for Evolutionary Anthropology. https://apics-online.info/contributions/46#tprimary

Summer Institute of Linguistics. (2021). *Glossary of Linguistic Terms.* https://glossary.sil.org/term/tense

Summer Institute of Linguistics. (2022). *Glossary of Linguistic Terms.* https: // glossary. sil.org /term/ clitic-grammar

Thomason, S. G. (2001). Language contact. Edinburgh University Press.

Velupillai, V. (2015). *Pidgins, creoles and mixed languages: An introduction.* Amsterdam: John Benjamins Publishing Company.

Viti, C. (2016). Synthetic and analytic structures for tense and aspect in Indo-european. In Bubenik, Vit & Drinka, Bridget & Hewson, John & Šefčik, Ondřej (eds.), *Exploring Universals of Tense and Aspect.*

Winford, D. (2005). Contact-induced changes. *Diachronica* 22 (2). 373–427.

Winford, D. (2006). *The restructuring of tense/aspect systems in creole formation*. In Ana Deumert, & Stephanie Durrleman (Eds.), Structure and Variation in Language Contact (pp. 85–110). Amsterdam: John Benjamins.

Wolff, J. U. (2006). *Encyclopedia of Language and Linguistics*. Boston: Elsevier.

Philippine English Proficiency of the K12 students: Basis for the improvement of the English Curriculum

Lemuel R. Fontillas, PhD
Bataan Peninsula
State University
lhemf@yahoo.com

Sherrilyn B. Quintos, EdD
Bataan Peninsula
State University
sbquintos@bpsu.edu.ph

Cynthia M. Ronquillo, MAEd.
Bataan Peninsula
State University
cmronquillo@bpsu.edu.ph

Abstract

This study attempted to assess the Philippine English proficiency of the students who took the K to12 curriculum. The study dealt with the speaking and comprehension skills of the students. The main aim of the study is to show the need to have proficiency in using the Philippine variety of English. The researchers interviewed students who were 18 to 19 years of age, male or female, and were enrolled for the first time in any of the programs offered by BPSU on the Main Campus. The recorded material exhibited each respondent's English comprehension and speaking skills; anonymity and privacy were observed. Recorded conversations were lexically transcribed hence forming a 336,828-word corpus. The analysis was done using 1) relevance of the answer, 2) eloquence, 3) grammatical content and 4) time of response. The study found that students can speak the language but use the Philippine English variety (Bautista, 2000a). Hence, we recommend that ELT teachers should have the acceptance of Philippine English. This can be done by having different awareness campaigns for the academic community. A curriculum modification for both the tertiary level and DepEd can be done to realign the use of Philippine English.

1 Introduction

In 2018, a new curriculum for the tertiary level was released by the Commission on Higher Education (CHED). The said curriculum was said to be more focused on the major subjects of the program. Many minor subjects such as Maths, Sciences, and languages were removed from the old curriculum and transferred to the Senior High School level. Language courses, particularly English, are not exempted from this. Specific courses such as Speech and Oral Communication, Writing in the Discipline, Study and Thinking Skills, Oral Diagnostic English and Communication Arts 1 & 2 were removed from the tertiary level. They were transferred to the new basic education program. These courses are avenues where our own variety of English can be introduced to students. As these courses were transferred to basic education, it is now the responsibility of the teachers to establish learning in the Philippine variety of English. This is despite the Philippine English (PE) paradigm still facing opposition from traditional teachers, or some teachers haven't entirely accepted the idea of having a Philippine variety of English at all Some teachers have the idea of having PE but are not confident enough to teach them to students. As the issue of having this variety is being taught in schools, it is evident that it already exists among Filipinos (Gonzalez, 1997). Teachers may have been teaching it unknowingly but lacking knowledge that they are dealing with PE already.

This study attempted to assess the Philippine English proficiency of the students who took the K to12 curriculum. The study dealt with the speaking and comprehension skills of the students. The main aim of the study is to show the need for proficiency in using the Philippine variety of English. Specifically, the study answered the following questions:

1. How is Philippine English displayed in the answers of the students in terms of their;
   1.1 relevance of the answer;
   1.2) eloquence;
   1.3) grammatical content; and
   1.4) time of response?
2. What is the implication of the developed skills of the students to language teaching?

2 Review of Literature

The language learning process of the Department of Education (DepEd) is anchored to a belief that for effective language acquisition and learning to take place, language teachers must be guided by the six (6) language teaching principles. These are Spiral Progression, Interaction, Integration, Learner-Centeredness, Contextualization, and Construction. These principles are applied in the classroom in which skills, grammatical items, structures, and various types of texts are taught, revised, and revisited at increasing levels of difficulty and sophistication. DepEd believes that this will allow students to progress from the foundational level to higher levels of language use. This may seem effective, but the National Achievement Test (NAT) 2018 showed declining scores and skills mismatch since 2013. This problem in the language education sector showed that despite continuous teaching of language skills, no mastery was taught to the student. The basic education ends at grade 12 then college is next. In College, where education is of a higher form, mastery of skills is enhanced but if mastery is underdeveloped what then will be enhanced?

In the United States, Carhill, Suárez-Orozco & Páez (2017) conducted a study to increase understanding of factors that account for academic English language proficiency in a sample of 274 adolescent first-generation immigrant students from China, the Dominican Republic, Haiti, Central America, and Mexico. Previous research has shown the importance of English language proficiency in predicting academic achievement measured by GPA and achievement tests. Their study described the academic English language proficiency of immigrant youth after, on average, 7 years in the United States and models factors that contributing to the variation. Findings show that although differences in individual student characteristics partially explain variation in English language proficiency, the schools that immigrant youth attended are also important. Students' time spent speaking English in informal social situations is predictive of English language proficiency. These findings demonstrate that social context factors directly affect language learning among adolescent immigrant youth and suggest a crucial role for school and peer interventions.

On the other hand, Pereda et al. (2013) pioneered a study about the influence of more than one language in second language acquisition through corpus analysis. Their study aimed to determine the possible interferences between a language and a target language. It was reported that the proposed project is innovative in that the conclusions are based on the Error Analysis of a large corpus and that the results were very useful for SFL teachers and learners in Flanders. At the same time, it fills an existing gap in SFL books and the topic of change-of-state verbs. It is argued that while the research clearly shows that many factors other than English proficiency are important to academic success, there may be for each institution, or even for each program within an institution, a minimum level below which lack of sufficient proficiency in English contributes significantly to lack of academic success. Such a level can be determined by each institution individually, but until it is determined, several steps can be taken to establish reasonable English language proficiency requirements.

Going back to the Philippines, where various English languages were born, Borlongan (2017) pointed out that in 1925, the educational survey board noticed that Filipinos spoke differently from Americans. Further contrastive reports (Raqueño, 1940, 1957) also pointed out the distinctive way of how Filipinos use English. Gonzalez (1997, 2008) said that when Filipino teachers began teaching fellow Filipinos English, which was around the 1920s, Philippine English was born, but it was only towards the end of the 1960s when a linguist, Teodoro Llamzon, called attention to an emerging variety of English in the Philippines.

Still, according to Borlongan (2017), through a publication of Llamzon in 1969, Philippine English had received much attention from Filipino linguists. It became an object of inquiry and was proven by some of the reviews in linguistic research in the Philippines done by Dayag and Dita (2012). Indeed, research on Philippine English had been remarked as the most likely comprehensive research among other indigenized Southeast Asian English (Tay 1991). Bautista (2000a) thus defined Philippine English: "Philippine English is not English that falls short of the norms of Standard American English; it is not badly learned English as a second language; its distinctive features are not errors committed by users who have not mastered the American standard. Instead, it is a nativized variety of English that has features that differentiate it from Standard American English because of the influence of the first language (specifically in pronunciation but occasionally in grammar), because of the different cultures in which the language is embedded (expressed in the lexicon and discourse conventions), and because of a restructuring of some grammar rules (manifested in the grammar)".

3 Methodology and Materials

3.1 Respondents of the Study

This descriptive qualitative study randomly selected first-year students from Bataan Peninsula State University (BPSU) Main Campus for the English Proficiency assessment. The university accepts an average of 2000 first-year students every semester, thus with a margin of error of 5% and a confidence level of 95%, a required sample size of 238 (students) is considered as respondents in this study. Respondents were 18 to 19 years of age upon the conduct of the interview of the teachers involved in the study, male or female, and were enrolled for the first time in any of the programs offered by BPSU on the Main Campus.

3.2 Corpus Building Process

The researchers of the study first discussed how to conduct the interview. To ensure the consistency and reliability of the data, the team had to select qualified respondents. The individual data recording took place in an area with less noise and fewer chances of interruption. The name and identities of the respondents were not asked to ensure anonymity and privacy. Participants were asked to answer a set of questions verbally. The answers of each respondent to the questions were audio-recorded; thus, it was just the interviewee's voice that was taped. Recordings were then transcribed for analysis. The respondents answered the following questions in English for at least four (4) minutes each: 1) Tell something about your experience in learning the English language. 2) How did you learn to use the English language? 3) Narrate unforgettable lessons your former English teachers taught you. 4) What are the topics in English that you have difficulty in mastering? Why do you think so? 5) What are you doing to make these difficulties ease?

3.3 Description of the Corpus

The recorded material exhibited the English comprehension and speaking skills of each respondent. All recorded conversations were lexically transcribed, forming a 336,828-word corpus for the study. The analysis was done using 1) relevance of the answer, 2) eloquence, 3) grammatical content, and 4) time of response. Based on the categories given, results were concluded. A speaking and comprehension rubric used in grading similar activities in purposive communication was used. The rubrics were prescribed in the region during the seminar on handling Purposive Communication; hence, these were validated prior to prescription.

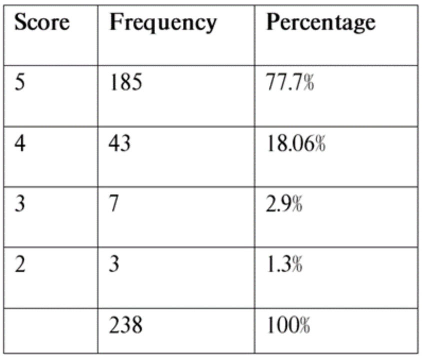
3.4 Corpus Management

Data were given codes to the mp3 files produced to ensure the organization and confidentiality of the respondents' answers. Alphanumeric codes were assigned such as S01 for Student 01, S02 for Student 02, and so on. The codes were the filenames of the mp3s and were used for the title of the transcribed material for analysis. The mp3s and the softcopy of the transcription were saved in a cloud application, Dropbox. The folder was password-protected to ensure the security of the files. A list of names of the respondents with their corresponding codes was also stored in the protected folder to secure the identity of the respondents as well. Only the researchers have complete access to the cloud folder.

4 Discussion of Results and Findings

4.1 Relevance of the Answer

Using the Comprehension Rubric, out of the 238 respondents, there were 185 students given a score of 5 for comprehension. This is 77.7% of the entire population. While 43 scored 4 and the remaining 10 scored 3 and 2. This minor number of respondents simply gave appropriate answers to questions, thus, expressed logically relevant statements. This shows that students can comprehend the interviewer's questions but have difficulty expressing their thoughts/ideas using the English language. To give a score of five (5) from the rubric means that the answer is given a full mark.



Score	Frequency	Percentage
5	185	77.7%
4	43	18.06%
3	7	2.9%
2	3	1.3%
	238	100%

Table 1. Score of Respondents

The description suggests that the initial post or answer is organized around a clear point of view or idea with adequate supporting detail. According to Gonzales (1982), Filipinos typically have mastery of the formal style or classroom English. Gonzales also concluded that there are minimal differences in the formal and informal written discourses. Loan words, nicknames, and contractions as often used in an informal style, and code-switching to the vernacular is generally prevalent in informal discourses. Respondents were bilingual speakers, and the corpus also revealed that Tagalog words appeared 781 times and a total of 1,426 wordy sentences, meaning the respondents knew what to say but they were gasping for the right equivalent word in the second language. De Boni (2006) explained that logical relevance could be based on and observed using measured simplification, a form of constraint relaxation, and considering flexibility and directness of statements in a sliding scale of aptness. This then shows that the respondents, despite their previous difficulties in learning the English language, are already familiar with it.

4.2 Eloquence

Table 2 refers to the respondents' coherence in answering the questions, while table 3 refers to their level of fluency and pronunciation of words. Using the speaking rubric, it was found out that 137 students out of the 238 respondents were coherent in their speaking. This is 57.56% of the population. According to the rubric, a 2.5 score means that a student correctly understands the questions and that responses are clear. Seventy eight (78) students or 32.77% got a score of 2, meaning that a student makes few mistakes understanding the questions, and responses are mostly clear.

Score	Frequency	Percentage
2.5	137	57.56
2	78	32.77
1.5	23	9.66
	238	100%

Table 2. Score of Respondents for Coherence

Lastly, 23 students, or 9.66%, got a 1.5 score, meaning that a student makes significant mistakes in understanding the questions, and responses are somewhat clear. The indicated scores show that majority of the respondents are coherent with their answers to the questions. This
is consistent with the result in table 1 on the relevance of the student's answers, which says 77.7% of the respondents scored 5

4.3 Fluency

In table 3, 125 or 52.5% of the respondents got a score of 2.5 which means students speak fluidly with few to no breaks. While 100 students, or 42.01%, got a score of 2, meaning students speak mostly fluidly with semi-frequent short or a few long breaks. Only 13 or 5.46% of the respondents got a score of 1.5, meaning these students speak somewhat fluidly with frequent short and long pauses, as seen from these numbers. It averages 40-60% of the respondents are eloquent in English.

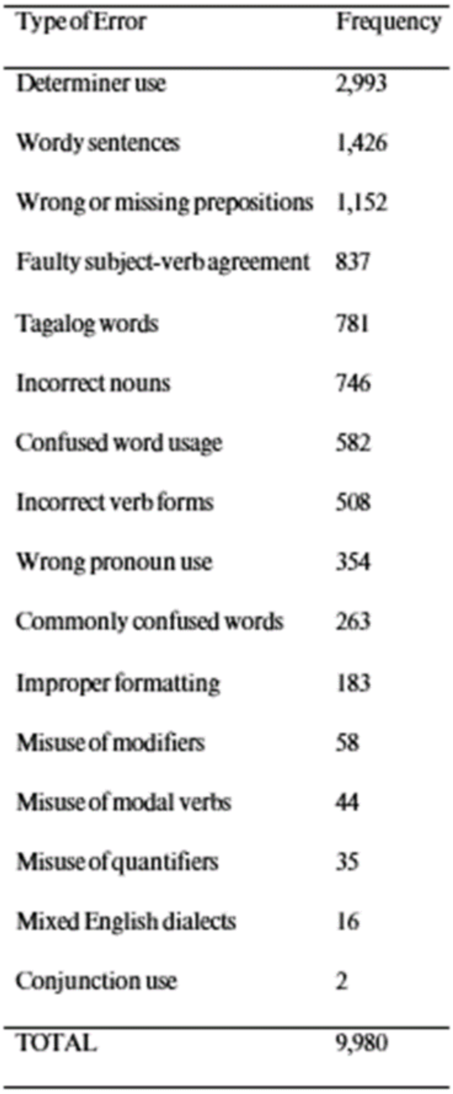
Score	Frequency	Percentage
2.5	125	52.5
2	100	42.01
1.5	13	5.46
	238	100%

Table 3. Score of Respondents for Fluency

The respondents were considered to be members of Gen Z; thus, there are influences on how they speak through the different media (both social and electronic) around them. It is also evident that their pronunciation of the words is unique in terms of sound.

It is similar to the tone of their speaking Filipino, their first language. This is proof of what Teodoro Llamson (1969) noted in his monograph, the distinction between Filipino and the American variety in producing vowel sounds, stress, and syllables. His study was expanded by Gonzales and Alberca (1978), who noted the distinctive features of Philippine English phonology as: absence of vowel reduction rule and possible spelling pronunciation, absence of schwa sound, the substitution of voiceless fricatives for voiced fricative, absence of aspiration of initial voiceless stops, simplification of consonant clusters, and different stress patterns in individual words, among others.

4.4 Grammatical Content



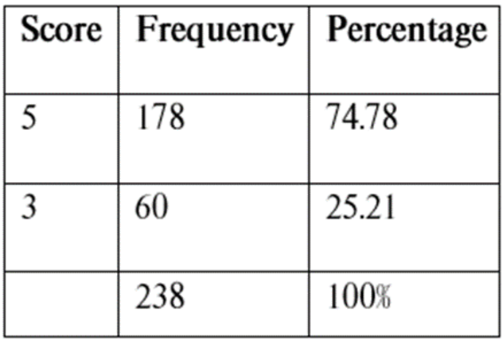
Type of Error	Frequency
Determiner use	2,993
Wordy sentences	1,426
Wrong or missing prepositions	1,152
Faulty subject-verb agreement	837
Tagalog words	781
Incorrect nouns	746
Confused word usage	582
Incorrect verb forms	508
Wrong pronoun use	354
Commonly confused words	263
Improper formatting	183
Misuse of modifiers	58
Misuse of modal verbs	44
Misuse of quantifiers	35
Mixed English dialects	16
Conjunction use	2
TOTAL	9,980

Table 4. Transcript Analysis of Errors

Table 4 reports a significant number of transcript errors associated with grammatical content, namely: determiner use (2,993), wrong or missing prepositions (1,426), Faulty subject-verb agreement (837), incorrect nouns (746), confusing word usage (582), incorrect verb forms (508), wrong pronoun use (354), improper formatting (183), misuse of modifiers (58), misuse of modal verbs (44), misuse of quantifiers (35), and conjunction use (2). All of these fall under intralingual interferences, which was defined by Erdogan (2005) as the errors resulting from the learners' view about the target language because of their lack of experience with it. These are the errors not related to the native language structure but caused by learners' limited target language information. These intralingual interferences manifested through the students' responses unveiled that there had been only partial learning of the target language. A corpus was formed out of the 238 respondents, having 336,828 words. The table below shows an error deviation of 9,980 words committed by the students. This proves what Gonzales and Alberca (1978) mentioned on the distinct variation in word order, article usage, noun subcategorization, as well as some errors in pronoun-antecedent agreement, tense -aspect usage, and subject-verb agreement. Bautista (2000a) noted similar findings in subject-verb agreement, articles, prepositions, mass and count nouns, word order, and comparative constructions. Instead of errors, Bautista adopted D'Souza's recommendation of categorizing variants that were rule-governed, widespread, and used by competent users as distinct features of Philippine English.

4.5 Time of response

Using the comprehension rubric, table 5 shows that 178 out of the 238, 74.78% of the total respondents, got a score of 5, and 60, 25.21%, scored 3. None got a score of zero since all have answers to all the questions. 74.78% answered immediately, while 25.21% of the respondents needed seconds to think for answers and probably got conscious of the recorder. Responding to questions is a normal reaction.



Score	Frequency	Percentage
5	178	74.78
3	60	25.21
	238	100%

Table 5. Score of Respondents
for Time of Response

The respondents know they are being interviewed and their answers are being recorded. However, their reaction to react quickly is an innate action coming from consciousness, but looking at what and how they answer the question still shows the evidence of having a variety of English uniquely embedded in them. The earlier variables discussed above show them.

5 Summary and Conclusion

Grammatical errors were committed by the respondents, as displayed in the data. As Bautista (2000) mentioned, common errors may be a nativized variety of English with different features from Standard American English. Thus, this only highlighted that the respondents in their age and level of English performance need to be guided to make a variety of English more intelligible without compromising the comprehensibility of the language. Language is a growing entity, and its changes are manifestations of it being alive. Thus, putting a stop to these changes and saying which is appropriate or not only hinders the growth, which leads to killing a language. Meanings are user and dependent. The English language is not exempted from this claim, just like any other language, it is a tool for expressing the thoughts, ideas, feeling, or simply the message the user wants to convey.

The respondents' so-called "grammatical errors" may not necessarily mean errors, as discussed above, but features of a language. These features then distinguish a particular variety giving birth to a language. These errors are most of the time related to how the respondents make use of their first language. On the other hand, errors not related to the native language structure are caused by the learners' limited information about the target language. These intralingual interferences manifested through the students' responses unveiled that there had been only partial learning of the target language. Thus, reinforcement is needed.

The analysis also revealed that there is also a need for coherence, fluency, and pronunciation of the respondents. Despite being members of Gen Z and being influenced by how they speak by the different media (both social and electronic) around them, there is still room for them to be proficient in using Philippine variety way of speaking. This can be proven when the students can comprehend the interviewer's questions but have difficulty expressing their thoughts/ideas using the English language. This proof strengthens the claim that the first language very much influences the English variety. Philippine English is evident in how the respondents/students performed in the study. Philippine English being a highly intelligible and acceptable language is dynamically expanding, and its rules and conventions in grammar, style, and usage is flexible and eclectic. Because of these characteristics, Philippine English is continually evolving, benefitting from a multi-dimensional effort of propagation through education, media, and literature.

6 Recommendations

The data that was gathered came from students enrolled before the pandemic. These students were a product of the K-12 curriculum thus, it is possible to have a new set of data coming from the HYFLEX setup with a new set of students and their answers to the same set of questions. Data from them to the new one can be compared to see if the improvements are already met in implementing the curriculum. Another area that can be investigated is that the learning modality at present is different thus, it would be beneficial to know for curriculum makers if learning can be adjusted for the betterment of the learners.

Based on the discussion above, it is also recommended that curriculum makers revisit the new General Education Curriculum for the tertiary level and the Senior High School curriculum by incorporating Philippine English as the variety to be taught can help students be proficient in using the variety. Aside from incorporating Philippine English in Purposive Communication and English Skills Enhancement at the College level, an additional course such as "Philippine English" will help enhance the evolving Philippine language learning. We cannot get away with proficiency. In the first place the study was conducted in an academic setting. The researchers themselves are academicians, who are responsible for making the students proficient in the field, in this case, in the English language macro skills field. Errors were identified as far as the other varieties are concerned, but there is a must that students be proficient in using Philippine English (PE). After all, PE is not a substandard variety of English. Curriculum makers can plan on making the curriculum more effective using the PE variety instead of other varieties. Teachers, on the other hand, can execute the learning to their students if they are also properly oriented with the World/ Philippine English paradigm. Kirkpatrick (2007) talks about how some scholars have recommended using a native speaker variety as a norm, with the local variety as a model. This being said, it is the heart of having a local variety such as Philippine English, where American English is the native speaker variety. In this way, learners will not be discouraged from using whatever variety they speak. It can be seen already in the present study that learners can talk about the language in the local variety. Imposing the local variety as part of the curriculum can produce Standard Philippine English.

Despite the awareness of the ELT teachers on the use of PE, acceptance of the topic is needed for other educators who do not advocate the use of this variety. For the new educators, symposiums, seminars, and awareness campaigns are suggested to be done. World Englishes varieties exist and being chauvinistic about a particular variety does not help make English language teaching more effective.

Relatively, more studies on using Philippine English in the academic setting are highly encouraged. This will further establish the emerging body of literature on the Filipino's own variety of English.

7 References

Andrew Gonzales. 1982. English is the Philippine mass media. New Englishes. Rowley, MA Newbury House Publishers, Inc.

Andrew Gonzales. 1997. The history of English in the Philippines. In M.L.S. Bautista (Ed.), English is an Asian language: The Philippine context-Proceedings of the conference held in Manila on August 2-3, 1996 (pp.25-40). North Ryde, Australia: The Maquaire Library Rty Ltd.

Andrew Gonzales. 2008. A favorable soil and climate: A transplanted language and literature. In M.L.S. Bautista & K. Bolton (Eds.), Philippine English: Linguistic and literary perspectives (pp.13-27). Hong Kong SAR, China: Hong Kong University Pres.

Andrew Gonzales & Wilfredo Alberca. 1978. Philippine English of the mass media, preliminary edition. Manila: De La Salle University Research Council.

Andy Kirkpatrick. (2007). Implications for international communication and English language teaching. Cambridge: CUP

Ariane M. Borlongan. 2017. Contemporary perspectives on Philippine English. The Philippine ESL Journal. 19. 1-9.

Carola Suárez-Orozco & Mariela M. Páez. 2017. Explaining English language proficiency among adolescent immigrant students. Educational Researcher, vol. 44, 3: pp. 151-160.

Danilo T. Dayag & Shirley N. Dita. 2012. Linguistic research in the Philippines: Trends, prospects and challenges. In V.A. Miralao & J.B. Agbisit (Eds.), Philippine social sciences: Capacities, directions, and challenges (pp.110-126). Quezon, the Philippines: Philippines Social Science Council.

Ma. Lourdes S. Bautista. 2000a. Defining Standard Philippine English: Its Status and Grammatical Features. Manila: De La Salle University Press.

Marco De Boni. 2006. Using logical relevance for question answering. Journal of Applied Logic, vol 5, 1: pp. 92-103. https://doi.org/10.1016/j.jal.2005.12.003

Mary W.J. Tay. 1991. Southeast Asia and Hong Kong. In J. Cheshire (Ed.), English around the world: Sociolinguistic perspectives (pp319-332). Cambridge, the United Kingdom: Cambridge University Press.

Noemi Pereda. et.al. 2013. Grammatical Change in the Verb Phrase in Contemporary Philippine English. Asiatic, Volume 10, Number 2.

Pedro G. Raqueño. 1940. A comparative study of the types of errors in English written compositions of Filipino high school students in 1926 and 1936 (Unpublished master’s thesis) . University of the Philippines, Quezon, the Philippines.

Teodoro A. Llamzon. 1969. Standard Filipino English. Philippines: Ateneo University Press.

Vacide Erdogan. 2005. Contribution of error analysis to foreign language teaching. Mersin Üniversitesi Eğiitim Fakültesi Dergisi, vol 1, 2: pp 261–270.

Verify: Breakthrough accuracy in the Urdu fake news detection using Text classification

Santosh
Institute of Business Administration, Karachi
Sahtiya68@gmail.com

Dr. Zarmeen Nasim
Institute of Business Administration, Karachi
znasim@iba.edu.pk

Toto
Institute of Business Administration, Karachi
toto.14879@khi.iba.edu.pk

Parkash
Institute of Business Administration, Karachi
parkash.14910@khi.iba.edu.pk

Abstract

Researchers around the world have been struggling to minimize the rising spread of fake news through several Natural Language Processing techniques and a great amount of work has been done for resource-rich languages like English, French, German, Spanish, Chinese, etc. Alternatively, minimal research has been carried out on the Urdu language, which is spoken by millions of people around the globe. This study works on solving the problem of detecting the authenticity of Urdu news through Text analytics and Natural language processing methods. Upon studying the previously conducted research on text analysis and classification in Urdu and other resource-rich languages, it was found that machine translation does not work very effectively for authenticity due to compromises in structure, grammatical accuracy, and vocabulary. Hence, during this study, a Text analytics model has been developed on the only publicly available Urdu news articles dataset, originally composed in Urdu and comprising 900 articles, 500 real and 400 fake. During the preprocessing, stop words, English words, characters and numbers, and punctuations were removed which affected negatively the accuracy of the model. Apart from that, the data was lemmatized and tokenized and their effects on judging the authenticity of the news articles were examined to be a positive development. The supervised learning models include Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), Decision Tree (DT), AdaBoost (AB), and XGBoost classifiers along with the combination of different sets of the word and character n-grams were applied to the data and their results were compared. As a result, it was found that the XGBoost classifier accompanied with the word unigram and 1-4 character n-grams generated 91% accuracy, the highest reported on this dataset so far.

Keywords: Fake news detection, Natural Language Processing, Media, Social Media, Urdu

1. Introduction

Fake news is defined as a factually incorrect news article, which aims at misguiding and misleading the readers. Researchers classify fake news into five categories (1) fabrication, (2) news satire, (3) manipulation (4) advertising (5) propaganda, and (6) news parody [4]. The events of the dissemination of fake news date back to the 15th century, centuries before the launch of mobile phones, social media, and the internet and have been reasons for many mishappenings among people of different countries, castes, communities, religions, and beliefs throughout the world.

With the development of digital and social media, information disseminates in seconds. People living in one corner of the world receive the information of another corner with just a click. This transformation in the methods of communication and speed of traveling of information has both pros and cons. Along with the spread of real information, it has been a medium for disseminating fake news quickly. Fake news has become a worldwide issue. People use it for several motives, like personal grudges, gains, and to change public opinion. It is perfectly evident from the scenario of the Covid-19 pandemic [2]. The circulation of fake stories about the victims, medical and homeopathic treatments, and the spread of the virus fueled panic among the people. WHO struggled to stop the dispersion of the fake Covid-19 stories related to the spread, symptoms, treatment, and prevention of the virus, but people in many countries followed the inbox forwarded methods and infected themselves with other diseases which led to panic in the public and disbelief in the competent authorities [2] [6].

For the prevention of dissemination of fake news, programmers and researchers apply National

Language Processing Techniques to identify the authenticity of news by identifying the characteristics, word choice, and writing patterns [1]. Extensive work has been done for high-resource languages like English because of the availability of the content, and literature [1][3]. Urdu, despite being spoken by more than 100 million speakers [1] around the world, is still considered a low-resource language because of the little literature and content available for Natural Language Processing Tasks [3]. This research aims at implementing Natural Language processing algorithms on Urdu scripted news and providing an open-source code as a resource for the confirmation of the authenticity of information for the Urdu-speaking public.

The unavailability of an authentic web source for the confirmation of the genuineness of the information in Urdu has caused several problems like Dr. Shahid Masood's imprisonment, and the Indian website's claim of a civil war in several cities of Pakistan. It has led to the arousal of conflicts within and outside and has deterred the country's image in national and international media [4]. Through this project, the researchers tend to solve the problems raised by the dissemination of false news and study the existing studies on fake news detection and NLP techniques applications in Urdu and other languages. This study examines the text classification techniques for confirming the trustworthiness of Urdu news articles, tries different sets of NLP techniques such as n-grams, lemmatization, etc, and compares their results, and in the result, it suggests the classification technique which serves the best in finding the genuineness of the content of an article of Urdu language.

2. Literature Review

There is little literature solely focused on fake news detection in the Urdu language. Hence, the researchers picked the commonly practiced NLP techniques for text processing and specifically for fake news detection in other renowned languages. In [1], for applying the NLP tasks, the data was cleaned by discarding the auxiliary character sequences and tokens, performing tokenization on words and characters, and the ramification of the sentences. For features, word n-grams from 1 to 6 and character n-grams from 1 to 6 were applied and the function words n-grams boosted their performances. A standard stop word list for Urdu was used as function words. For the binary classification of the news articles in [1], several methods were implemented which include Multinomial I Bayes, BernoINaive Bayes, Support Vector Machines, Logistic Regression, Random Forests, Decision Tree, and AdaBoost. The AdaBoost lent the maximum F1 score with particular combinations of character-word 2-grams and unigrams.

The team working on the study [1] continued their work in the study [3]. In this study, the dataset was enhanced by adding 400 news articles that were originally in the English Language and were machine translated into the Urdu Language by Google translate. Combined, the dataset contained 700 real news and 600 fake news articles, out of which 900 articles were used in the study [1]. The best classifiers from the study [1], SVM and AdaBoost were applied and it was found that the experimentation on the augmented dataset performed lesser than the original Urdu news articles because of the imperfect quality of the machine translation, which was also confirmed manually.

In 2020, the Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico conducted a fake news detection challenge for the Urdu Language, in which 39 teams participated. The teams had to perform the task with maximum accuracy on the dataset used in [1]. The team for [5] applied the generalized autoregressors technique for the binary classification task. They trained the XLNet model that uses the AR pre-training method and employs the use of language modeling objectives based on permutation. Their system reported an overall accuracy of 0.84 and an overall F1 macro score of 0.83.

The team from [1] and [3] also participated in this competition with the name BERT 4 EVER and stood first with the highest accuracy which led to the study [4]. In this task, the dataset was increased and more news articles were collected in the same way as done in the already available dataset from the study [1], adding 400 news articles, and bringing the test dataset to a total of 400 news articles and the train dataset to 900 news articles.

Furthermore, quoting the techniques used by the teams it must be mentioned that only one of the teams removed stop words during the preprocessing while other teams did not remove the stop words from the data, the teams used different techniques of text representation. Three of them used weighted tf-idf, three of them represented the articles using word embeddings while two teams applied different approaches of Word2Vec and FastText embeddings respectively. BERT4EVER used the contextual representation using BERT, which is a recent and advanced manner of text representation. To classify the corpus some teams used the classical non-neural algorithms while others' submissions consisted of various neural network architectures. Among the submitted models, the team BERT 4 EVER outperformed the character bi-grams with logistic regression baseline achieving an F1-macro score of 0.90. This fact confirms that contextual representation and large neural network techniques perform better than the classical feature-based models [4]. It has also been shown in many recent studies in all branches of natural language processing.

Due to the unavailability of the originally labeled datasets in Urdu, the study [7] was conducted on English-translated news articles. The study was conducted on a translated QProp English language dataset containing 5,322 fake and 6,252 real news articles. On manual confirmation, it was found that around 95.4% of the translation was accurate. They introduce Propaganda Spotting in Online Urdu Language –ProSOUL) - a framework to identify content and sources of propaganda spread in the Urdu language. The team developed a Linguistic Inquiry and Word Count dictionary for the extraction of psycho-linguistic features of the Urdu Language. For the text representation, n-gram, NELA, word2Vec, and BERT features and the combination of word n-gram, character n-gram, and NELA features led to the best performance with 0.91 accuracies. In the comparison of the BERT features, Word2Vec performed better than BERT technique for the word embeddings [7].

In the studies [8], and [9] data mining techniques have been used for the detection of fake news by classifying the posts, and online reviews in publicly available corpora. The research team working on [9] achieved 99.7% accuracy by using a logistic classifier implemented in the browsers of the users. In [10], a rumor verification model has been proposed that achieves improved performance for veracity classification by leveraging task relatedness with auxiliary tasks, specifically rumor detection and stance classification, through a multi-task learning approach.

Similarly, the studies [10],[11],[12] have focused on the feature extraction from the text and then used those features in the classification models that include Logistic Regression, K-Nearest Neighbor, Random Forest, and Support Vector Machines. Study [10] reports on comparative style analysis of hyperpartisan (extremely one-sided) news and fake news. This study shows how a style analysis can distinguish hyperpartisan news from the mainstream (F1 = 0.78), and satire from both (F1 = 0.81). In [11], the researchers present a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics, and representative datasets.

Several studies have been conducted to find the relationship between the title and the content instead of classifying them into real or fake [13][14]. In the study [13], 73% accuracy was achieved, which was 26% higher than the previously conducted research by Excitement Open Platform[13]. For the study [14], the proposed approach is based on a maximum entropy classifier, which uses surface-level, sentiment, and domain-specific features represented in the Tweet Stance Detection task in SemEval 2016.

For this study, all these previously conducted fake news detection studies in Urdu and other languages were studied and the techniques were adopted for experimentation on the Urdu dataset used in this research.

3. Research Methodology

As displayed in figure 1, the workflow of the project was divided into several steps which include Dataset Collection, Data preprocessing, modeling, and evaluation. There were multiple substeps under these steps which further elaborated the diversity of the task and opened more doors to possible experimentations for further research. The classifiers and techniques were evaluated using accuracy, F1 real, and F1 Fake measures, and out of them, accuracy (Test score) was the deciding factor for the efficiency of a specific algorithm. Each step shown in the workflow is described as follows:

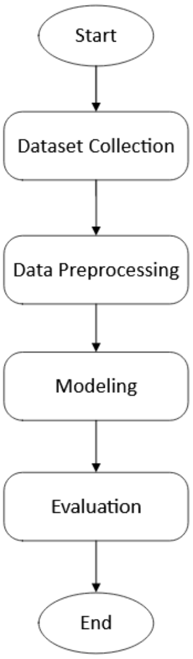


Fig. 1. Stepwise research methodology

3.1 Dataset Collection

The dataset for this study has been adopted completely from GitHub[1], originally posted by the team working on the study[1]. As mentioned above, it is the only labeled dataset available in the Urdu language, which helps to run the techniques and get the results with maximum efficiency, as translating the articles from another language deteriorates the performance. For the collection of the real news, the researchers of the study [1] crawled the data from trustworthy websites like BBC Urdu, Geo News, etc, or by verifying the news on multiple platforms. They crawled the articles using the Newspaper library of Python[2] which provides ease in dealing with the HTML, noisy texts, images, and advertisements on different web pages.

On the other hand, for gathering the fake news, professionals, and journalists were hired to write the fake news articles manually during the study [1]. It saved the team a great hustle of searching and verifying the fake news on online platforms. The journalists were directed to imitate the real news writing pattern so that there are no clear indications for the articles to be judged their authenticity easily[1].

The dataset contains a total of 900 news articles, 500 labeled real and 400, fake. The data has been divided into the train and test datasets by 638 and 262 articles respectively, where the train part is comprised of 350 real and 288 fake articles, and the test set is comprised of 150 real and 112 fake articles. The news in these articles belongs to five categories, (i) Business, (ii) Health, (iii) Showbiz,(iv) Sports, and (v) Technology. The category-wise distribution of the articles is displayed in Table 1.

Category	Train		Test	
	Real	Fake	Real	Fake
Business	70	36	30	14
Health	70	70	30	30
Showbiz	70	70	30	30
Sports	70	42	30	8
Technology	70	70	30	30

Table 1. Dataset distribution by category of news articles

3.2 Data Preprocessing

Before starting the classification experiments, several preprocessing and data quality-enhancing experiments were conducted to help increase the accuracy of the model. The following four tasks were implied.

1) Removing the English and Urdu Punctuation
2) Removing English and Urdu Numbers
3) Removing English words and characters
4) Removing stop words

These tasks were implemented individually as well as combined and the effect was studied.

3.3 Modeling

This section of the study discusses the experiments conducted on the dataset for solving the problem of identifying fake news by applying the various sets of classification algorithms along with different sets of character and word n-grams. Upon applying lemmatization and tokenization, the effects on the accuracy of all seven (7) classification models were studied. Each supervised learning model was tested against each set of character and word n-grams, and the results were compiled. Finally, the techniques which helped to reach the goal of enhancing the evaluation measures were used in the final classification model.

[1] https://github.com/MaazAmjad/Datasets-for-Urdu-news

[2] https://newspaper.readthedocs.io/en/latest/

3.3.1 Lemmatization

Lemmatization is a technique that reduces inflectional forms and sometimes derivationally related forms of a word to a common base form. It is carried out by using the vocabulary of a particular language and it returns the dictionary form of a word which is known as a lemma. For instance, a lemmatization process reduces the inflections, "am", "are", and "is", to the base form, "be". Sometimes the concept is interrelated or misunderstood with stemming, which is used to collapse the derivationally related words. Lemmatization on the datasets was performed by using the Stanza library of Python[3] which supports multiple languages including Urdu. Applying this technique helped the model to work efficiently, improve accuracy, and fester accurate results by decreasing the noise of the data, and using the contexts of the words in the dictionary form.

3.3.2 Tokenization (n-grams)

An n-gram is the sequence of n-items. In other words, it is the combination of adjacent words where n represents the number of items for example unigram represents 1 item, bigram represents 2 items, and so on. In this study, N-gram features are used to build fake news detection models such as character n-grams, and word n-grams. These character and word n-grams divide the words into characters and sentences u into words respectively, which help the model identify the trend and likeability of the proceeding word or characters. This way, the model of Natural Language Processing can identify the upcoming characters or words and their correspondence to specific positive or negative results. In other words, by using tokenization, the model can be trained on sets of corresponding characters and words to better identify or predict the similarities for test datasets, which leads to better identification of the subclass of the under-test item.

3.3.2.1 Word n-grams

Word n-grams represent the n number of words sequence, for example, consider the sentence “I am a Data Scientist”. For word unigram, it will be divided into, “I”, “am”, “a”, “Data”, and “Scientist”. During the experimentations, 1 and 2 sequences of words, namely word unigrams and word bigrams on the data were implemented and the results were examined for all the supervised learning algorithms.

3.3.2.2 Character n-grams

Similar to the word n-grams, the character n-grams represent the number of character sequences, for example, for the word “Data”, the character unigrams would be “D”, “a”, “t”, and “a”. In this study, researchers used 1,2,3,4,5,6-character n-grams individually as well as merged. These character n-grams were then tested against each classification algorithm accompanied by the word n-grams. The combination of different words n-grams and characters n-grams helped to get deeper insights into the data and suitable combinations for accomplishing the goal.

3.3.3 Classification models

In this research, multiple classification models and techniques were considered and their performance on the given word n-grams and character n-grams were examined for the detection of the authenticity of the news [1]. The classifiers include Multinomial Naive Bayes (MNB) [16], Bernoulli Naive Bayes (BNB) [16], Support Vector Machines (SVM) [17], Logistic Regression (LR) [21], Random Forests (RF) [22], Decision Tree (DT) [24], AdaBoost (AB), and XGBoost Classifier. The models have been described briefly as under.

3.3.3.1 Multinomial Naive Bayes (MNB)

It is a widely used classification model. Given a set of labeled data, the model often uses a parameter learning method called Frequency Estimate (FE), which computes appropriate frequencies from the data and calculates the probabilities of the words. The model is efficient for text classification and easy to implement.

3.3.3.2 Support Vector Machines (SVM)

An SVM classifier creates a maximum margin hyperplane that lies in transformed input space and splits the example classes while maximizing the distance to the nearest cleanly split examples. The parameters of the solution hyperplane are derived from a quadratic programming optimization problem [17].

[3] https://stanfordnlp.github.io/stanza/

3.3.3.3 Logistic Regression (LR)

Logistic regression is a predictive analysis like all regression analyses having the dichotomous (binary) dependent variable. It is used to describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables [21]. It is more like a linear regression with added complexities of the cost function, which is known as sigmoid or logistic function.

3.3.3.4 Decision Tree (DT)

It is a supervised machine learning algorithm, which is used to classify the given record or to predict the outcome of regression problem. Generally, features in the data are placed on the non-leaf nodes while the branches contain the decision criteria. Every leaf node is a possible outcome of the problem [23]. Decision tree analysis is a divide-and-conquer approach to classification. They can be used to discover features and extract patterns in large databases that are important for discrimination and predictive modeling. Decision trees have an established foundation in both the machine learning and artificial intelligence literature and are slowly developing a niche in both the chemical and biochemical sciences [24].

3.3.3.5 Random Forests (RF)

Random forest is an ensemble of classifying and predictive machine learning algorithms used to solve more complex problems. It is a forest of many decision trees using bagging or bootstrap methods of aggregation. Random decision forests easily adapt to nonlinearities found in the data and therefore tend to predict better than linear regression. More specifically, ensemble learning algorithms like random forests are well suited for medium to large datasets.

3.3.3.6 AdaBoost (AB)

It is a boosting technique of machine learning which follows the sequence apply, boosts the previously learned model, and adapts. It repetitively follows this sequence and gets better and better results each time.

3.3.3.7 XGBoost Classifier.

XG Boost is a boosted version of the gradient boosting framework machine learning algorithm which is based on the decision tree. Each tree in the XG boost model boosts attributes that lead to the misclassification of the previous tree. The flexibility and speed of this technique provide it an edge over many algorithms in terms of efficiency, and validity.

3.3.4 Vectorizer

The tf-idf vectorizer was implemented on the data. A tf-idf vectorizer is a widely accepted technique for text vectorization, for Bag of Words representation. Each document is represented as a vector and the terms in the document are represented by the fields. In tf-idf, tf represents the number of times a term has appeared in a document and the idf denotes how informative the term is. The higher the repetition of a term across the documents the lesser informative it is considered to be. The tf-idf of a term is calculated as

$$TF.IDF = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad Eq.1$$

Where $tf_{\,i,\,j}$ = Number of occurrences of i in document j

$df_{\,i}$ = Number of documents containing i
N = Total number of documents

3.4 Evaluation

As mentioned previously, multiple classification algorithms along with the combination of word and character n-grams were applied to the data, and their effects were studied. Test score (Accuracy), F1 Real, F1 Fake, Precision, and recall. Out of these parameters, accuracy (Test Score) was the main metric that denotes the efficiency of the classifier along with the combination of the word and character n-grams. As the previously conducted studies on fake news detection also judge the efficacy of the algorithms based on the test score, hence the results of the proposed approach can be compared to them.

The F1 score is called the harmonic mean of precision and recall. Precision and recall are metrics of performance more suitable for imbalanced data because they allow taking into account the type of errors (false positives or false negatives) that the model makes

The accuracy score can be calculated by:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \qquad Eq.2$$

4. Result Analysis

This section tends to describe the results of all the experimented tools and techniques, and their effect on the accuracy of the model. More than 30 test experiments were performed which catered to almost all the combinations of preprocessing, word and character n-grams, and classified learning algorithms. For each step, the efficacy on the model was compiled, and the next approach was decided based on the previous results. All the experimentations carried out on the data are explained as under.

4.1. Preprocessing

From the results of preprocessing, it could be found that removing the punctuation, English and Urdu numbers, the English words and characters, and stop words would be a bad decision as it led to lesser accuracy in the experiments. Further, after extracting the features, it was found that the fake to real news punctuation count ratio was around 1:2, as displayed in Table 2.

Target	Word Count	Unique word	Stop words	Mean word Length	Character Count	Punctuation count	English character count
0	355.66	178.42	174.24	3.74	1682.39	16.54	8.37
1	298.8	155.51	150.91	3.62	1374.28	8.9	1.55

Table 2. The features extraction of the real and fake news articles (Target = 0 (Real news), Target = 1 (Fake News)

The features extracted from the data proved to be very helpful in recognizing the authenticity of the news for example the significant difference in the punctuation count and English characters count, as in fig 2. Also, the number of stop words in the real news articles was around 17.1% more than that of the fake news articles. Removing the punctuations or these characters would have blurred the difference between the two.

4.2 Lemmatization

After performing the lemmatization, the results were studied by checking the accuracy of multiple classifiers, and it was found that lemmatization affected negatively the accuracy of the models. So, after the confirmation, the dataset was not lemmatized for further evaluation as it would have deteriorated the model's efficiency to find out the authenticity of the articles in the dataset.

4.3 Tokenization

Classifier	Word Grams	Char Grams	Total Features	Accuracy (TestScore)	F1 Real	F1 Fake
AdaBoost	0	2,3,4,5	317276	0.83	0.86	0.78
RandomForest	0	2,3,4,5	317276	0.82	0.86	0.75
RandomForest	1	2, 3, 4, 5	327987	0.81	0.85	0.71
AdaBoost	1, 2	3, 4	199709	0.79	0.82	0.74
AdaBoost	1	3, 4, 5	327987	0.78	0.82	0.72
AdaBoost	1	2, 3, 4, 5	327987	0.77	0.8	0.72
RandomForest	1	3, 4, 5	327987	0.74	0.8	0.62
AdaBoost	1	0	11569	0.74	0.78	0.66
RandomForest	1, 2	3, 4	199709	0.74	0.8	0.62
RandomForest	1	0	11569	0.72	0.79	0.56
RandomForest	1, 2	0	78712	0.72	0.79	0.57
AdaBoost	1, 2	0	78712	0.71	0.77	0.62
NaiveBayes	1	0	11569	0.59	0.74	0.03
NaiveBayes	1	3, 4, 5	327987	0.58	0.73	0.00
NaiveBayes	0	2,3,4,5	317276	0.58	0.73	0.00
NaiveBayes	1	2, 3, 4, 5	327987	0.58	0.73	0.00
NaiveBayes	1, 2	0	78712	0.58	0.73	0.00
NaiveBayes	1, 2	3, 4	199709	0.58	0.73	0.00

Table 3. Classification models and their performances based on different sets of word n-grams and character n-grams

Different sets of character and word n-grams were applied belonging to each binary classification algorithm. From the experiments, it was found that applying tokenization improved the model's capability to confirm the real/fake nature of the given part of the data, and was counted as a positive development, as evident in Table 3.

4.4 Classification

The research found that the boosting classification algorithms performed the best as compared to other classifiers, with the XGBoost classifier outperforming all the classifiers with a combination of 1,2,3 word n-grams and 1,2,3,4 character n-grams and an accuracy of 91%, F1 Real Score and F1 fake score equals to 0.93 and 0.89 respectively. These scores are the highest scores achieved so far. AdaBoost classifier got the second position in achieving the goal. Multiple character n-gram and word n-gram were implemented on the XGBoost classifier. The top five best-performing word n-gram and character n-gram features combination are represented in table 4.

Classifier	**Word Grams**	**Char Grams**	**Total Features**	**Accuracy (Test Score)**	**F1 Real**	**F1 Fake**
XGB	1,2,3	1,2,3,4	376318	0.91	0.93	0.89
XGB	1	1,2,3,4	151945	0.89	0.9	0.85
XGB	0	1,2,3	19756	0.8	0.84	0.74
XGB	1,2	1,2,3,4	181254	0.8	0.83	0.75
XGB	0	1,2	1539	0.79	0.83	0.72

Table 4. Word n-grams, character n-grams, and the relevant score in metrics for XGBoost Classifier

5. Conclusion and Future Work

Fake news is a major problem in today's world. The detection of fake news is a promising task for all languages to help people be aware of the truth. Fake news dispersal is a serious issue that needs to be addressed and tackled in many languages. Researchers have proposed different data mining techniques for detection of the fake news in multiple international languages. But, there is a large room for improvement and creativity in the Urdu language which is spoken by a large community. The availability of very few previous studies for fake news detection in the Urdu language is a major limitation in conducting this research. Hence this study also adopts the methods and techniques previously applied for fake news detection in other renowned resource-rich languages.

This research contributed to filling this gap to the best possibilities by classifying the news articles of the first-ever labeled dataset in the Urdu language. The study applied Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), Decision Tree (DT), AdaBoost (AB), and XGBoost classifiers along with the combination of different sets of the word and character n-grams. The proposed approach in this study achieved 91% accuracy, the best accuracy with the highest scores so far using the XGBoost classifier with the combination of word unigram and 1-4 character n-grams.

This study attempts to fulfill the gap in Urdu fake news detection to the best possibilities and create room for further research on feature extraction techniques and classifiers. The boosting algorithms and n-grams used in this study can be researched further to enhance the accuracy of the model. Apart from this, the enhancement of the dataset can be a major add-on that will help to deeply learn the characteristics and the writing patterns of authentic and unauthentic writings. This way, it will be easier to distinguish between the truth and the lie.

References

Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., & Gelbukh, A. (2020). "Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation. Journal of Intelligent & Fuzzy Systems, 1–13. doi:10.3233/jifs-179905

Destiny Apuke, O., & Omar, B. (2020). Fake News and COVID-19: Modelling the Predictors of Fake News Sharing Among Social Media Users. Telematics and Informatics, 101475. doi:10.1016/j.tele.2020.101475

Amjad, M., Sidorov, G., & Zhila, A. (2020). Data Augmentation using Machine Translation for Fake News Detection inthe Urdu Language. Proceedings of the 12th Conference on Language Resources and Evaluation.

Amjad, M., Sidorov, G., Zhila, A., Gelbukh, A., & Rosso, P. (2020). Overview of the Shared Task on Fake NewsDetection in Urdu at FIRE 2020. Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico.

Khiljia, A. F., Laskara, S. R., Pakraya, P., & Bandyopadhyaya, S. (2020). Urdu Fake News Detection using Generalized Autoregressors. aDepartment of Computer Science and Engineering, National Institute of Technology Silchar, Assam, India.

Jahangir, R. (2020, March 28). Dawn. Desi totkas and fake news — a guide to surviving the Covid-19 'infodemic' . Dawn. Retrieved from Dawn.com: https://www.dawn.com/news/1544256/desi-totkas-and-fake-news-a-guide-to-surviving-the-covid-19-infodemic

KAUSAR, S., TAHIR, B., & MEHMOOD, M. ,. (2020). ProSOUL: A Framework to Identify PropagandaFrom Online Urdu Content. IEEE Access.

V. P´erez-Rosas, B. Kleinberg, A. Lefevre and R. Mihalcea, Automatic Detection of Fake News, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, (2018), pp. 3391–3401.772https://www.aclweb.org/anthology/C18-1287.773[8]

M. Aldwairi and A. Alwahedi, Detecting Fake News in Social Media Networks,Procedia Computer Science141(2018), 215–222. https://linkinghub.elsevier.776com/retrieve/pii/S1877050918318210.

M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff and B. Stein, A Stylometric Inquiry into Hyperpartisan and Fake News, in:Proceedings of the 56th. Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, (2018), pp. 231–240.

K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, Fake News Detection on Social Media: A Data Mining Perspec-tive, ACM SIGKDD Explorations Newsletter19(1) (2017), 22–36.

J.P. Posadas-Dur´an, H. G´omez-Adorno, G. Sidorov and J. Jaime Moreno Escobar, Detection of Fake News in a New Corpus for the Spanish Language, Journal of Intelligent & Fuzzy Systems(2018).

W. Ferreira and A. Vlachos, Emergent: a Novel Data-Set for Stance Classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL' 2016, (2016), pp. 1163–1168.804.

P. Krejzl, B. Hourov´a and J. Steinberger, Stance Detection in Online Discussions,arXiv preprint arXiv:1701.00504806(2017)

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), 76–81. doi:10.1109/mis.2019.2899143

Wang, S., Jiang, L. & Li, C. Adapting naive Bayes tree for text classification. Knowl Inf Syst 44, 77–89 (2015). https://doi.org/10.1007/s10115-014-0746-y

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and Their Applications, 13(4), 18–28. doi:10.1109/5254.708428

Shmilovici, A. (2009). Support Vector Machines. Data Mining and Knowledge Discovery Handbook, 231–247. doi:10.1007/978-0-387-09823-4_12

Dey, A., Jenamani, M., & Thakkar, J. J. (2017). Lexical TF-IDF: An n-gram Feature Space for Cross-Domain Classification of Sentiment Reviews. Pattern Recognition and Machine Intelligence, 380–386. doi:10.1007/978-3-319-69900-4_4

Bhattacharjee, U., P.K, S., & Desarkar, M. S. (2019). Term Specific TF-IDF Boosting for Detection of Rumours in Social Networks. 2019 11th International Conference on Communication Systems & Networks (COMSNETS). doi:10.1109/comsnets.2019.8711

Chao-Ying Joanne Peng , Kuk Lida Lee & Gary M. Ingersoll (2002) An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, 96:1, 3-14, DOI: 10.1080/00220670209598786

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal: Promoting Communications on Statistics and Stata, 20(1), 3–29. doi:10.1177/1536867x20909688

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660–674. doi:10.1109/21.97458

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics, 18(6), 275–285. doi:10.1002/cem.

Extractive Text Summarization with Latent Topics using Heterogeneous Graph Neural Network

Tuan-Anh Phan, Ngoc-Dung Nguyen, and Khac-Hoai Nam Bui *
Viettel Cyperspace Center, Viettel Group
Hanoi, Vietnam
{anhpt161,dungnn7,nambkh}@viettel.com.vn

Abstract

This paper presents a heterogeneous graph neural network (HeterGNN) model for extractive text summarization (ETS) by using latent topics to capture the important content of input documents. Specifically, topical information has been widely used as global information for sentence selection. However, most of the recent approaches use neural models, which lead the training models more complex and difficult for extensibility. In this regard, this study presents a novel graph-based ETS by adding a new node of latent topics into HeterGNN for summarization (TopicHeterGraphSum). Specifically, TopicHeterGraphSum includes three types of semantic nodes (i.e., topic-word-sentence) in order to enrich the cross-sentence relations for extractive summarization. Furthermore, an extended version of TopicHeterGraphSum for multi documents extraction is also taken into account to emphasize the advantage of the proposed method. Experiments on benchmark datasets such as CNN/DailyMail and Multi-News show the promising results of our method compared with state-of-the-art models.

1 Introduction

ETS is an important task of Natural Language Processing (NLP) in terms of extracting several relevant sentences from the original documents while keeping the main information. The traditional methods for ETS are TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004), which focus on calculating the similarity between sentence scores. Sequentially, the rapid development of Deep Learning (DL) has brought breakthrough records by modeling a document as a sequence of sequences in order to deal with long-range inter-sentence relationships for the summarization (Cheng and Lapata, 2016; Cohan et al., 2018). However, cross-sentence relation is still a challenge in this research field (Liu and Lapata, 2019). Recent works focus on Graph Neural Networks (GNNs) (e.g., Graph Convolutional Network (GCN) (Kipf and Welling, 2017) or Graph Attention Network (GAT) (Velickovic et al., 2018)) to explore the cross-sentence relationships for the summarization task. The core idea is to represent inter-sentential graphs and use message passing to extract the complex relationship in the input documents. For instance, (Yasunaga et al., 2017) and (Xu et al., 2020) adopt discourse analysis to build document graphs. (Jia et al., 2020) and (Wang et al., 2020) built a bi-partite graph between words and sentences, which is regarded as a heterogeneous graph neural network. Moreover, modeling global information is also taken into account for sentence selection by using pre-trained models (Liu and Lapata, 2019; Zhang et al., 2019).

Sequentially, (Cui et al., 2020) utilized pretrained BERT to learn contextual sentence representations and train jointly with latent topics using the neural topic model (NTM). (Nguyen et al., 2021) presents an extended version using NTM for abstractive text summarization indicating the capability of enriching the global information for the summarization.

*Corresponding Author

Although the existing methods have provided remarkable results, there are several open research issues that need to take into account: i) the high performance mainly depends on pre-trained models for learning sentence representations, which is difficult for the extensibility, especially for low-resource languages; ii) the current external information (e.g, latent topics) are extracted by neural models, which requires more complex configurations of the training process. Furthermore, the model might be suffering because of the bias problem, especially in terms of small datasets; iii) multi-document summarization is still an open research issue, which requires a comprehensive summary for covering an event and avoiding redundancy. In this regard, this study proposes a new HeterGNN model for the ETS problem by adding latent topic nodes into a graph structure, in which the initialized topic features are extracted by well-known clustering methods such as K-means and Gaussian Mixture Models (GMM). The core idea is to investigate the impact of topical information on the EDS problem in terms of both single and multiple document extraction. To the best of our knowledge, this paper is the first study to adopt topical information for multi documents summarization by using the concept of the heterogeneous graph structure. More detail of the proposed model is described in the following sections.

2 Background

The proposed model is based on the concept of a HeterGNN model, which is proposed by (Wang et al., 2020), for enriching the relationships between sentences by adding nodes with semantic features. Specifically, Fig. 1 illustrates the HeterGNN for the ETS problem. Particularly, the model includes three main components, such as initialized graph structure, graph layer, and sentence selection module. Graph structure is initialized by the set of word nodes, which is encoded using Glove (Pennington et al., 2014) as the addition node, and sentence features, which are calculated by combining CNN for extracting the local n-gram feature of each sentence and bidirectional Long Short-Term Memory (BiLSTM) for extracting the sentence-level feature, respectively. In this regard, the feature of the sentence s_j can be obtained as follows:

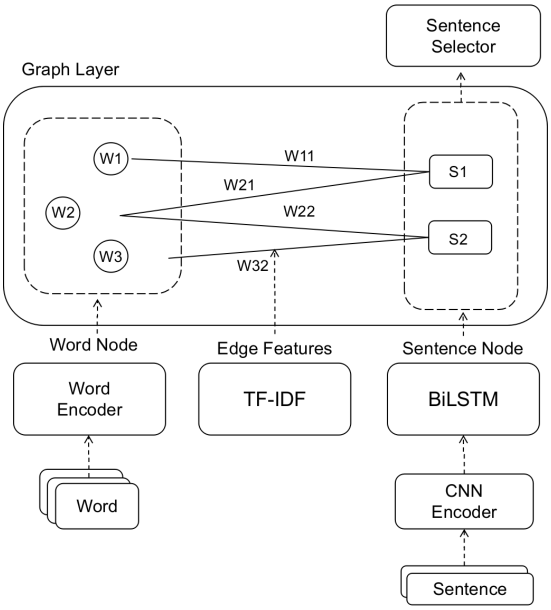


Figure 1: Model overview of HeterSumGraph

$$X_{s_j} = CNN(x_{1:p}) \oplus BiLSTM(x_{1:p}) \quad (1)$$

where p denotes the number of words in the sentence. Moreover, tf-idf is adopted for further approval of information between words and sentences. Sequentially, the graph layer is updated using GAT(Velickovic et al., 2018), with a modification for the heterogeneous graph. Specifically, the updated node representation with modified GAT can be formulated as follows:

$$z_{ij} = LeakyReLU(W_a[W_q h_i; W_e h_j; \overline{e}_{ij}]) \quad (2)$$

where $\overline{e}_{ij}$ denotes the multi-dimensional embedding space ($\overline{e}_{ij} \in R^{d_e}$), which is mapped from edge weight e_{ij}. Thereby, the sentences with their neighbor word nodes are updated via modified-GAT and Position-Wise Feed-Forward (FFN) layer, which can be sequentially formulated as follows:

$$\begin{aligned} U^1_{s\leftarrow w} &= GAT(H^0_s, H^0_w, H^0_w) \\ H^1_s &= FFN(U^1_{s\leftarrow w} + H^0_s) \end{aligned} \quad (3)$$

where H^0_w and H^0_s are the node features of word X_w ($X_w \in R^{m\times d_w}$) and sentences X_s ($X_s \in R^{n\times d_s}$), respectively. Therefore, the new representations of word nodes can be obtained using the updated sentence nodes and further updated sentences or query nodes, iteratively. Each iteration contains a sentence-to-word and a word-to-sentence update process, which can be demonstrated as follows:

$$\begin{aligned} U^{t+1}_{w\leftarrow s} &= GAT(H^t_w, H^t_s, H^t_s) \\ H^{t+1}_w &= FFN(U^{t+1}_{w\leftarrow s} + H^t_w) \\ U^{t+1}_{s\leftarrow w} &= GAT(H^t_s, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_s &= FFN(U^{t+1}_{s\leftarrow w} + H^t_s) \end{aligned} \tag{4}$$

The output of the new sentence representation is input into a sentence classier, which uses cross-entropy loss, for ranking the classification.

3 Methodology

In this study, our model is proposed for single document summarization (SDS), however, it can be extended for multi documents (MDS) with minor modifications. The methods for the two aforementioned problems are described in the following sections.

3.1 Single Document Summarization

Given an arbitrary document $d = \{s_1, .., s_n\}$, which includes n sentences, the objective of EDS for single document problem is to predict a set of binary label $\{y_1, .., y_n\}$ ($y_j \in [0, 1]$), which determine that the sentence in the summary or not. Figure 2 illustrates the structure of the proposed HeterGNN model.

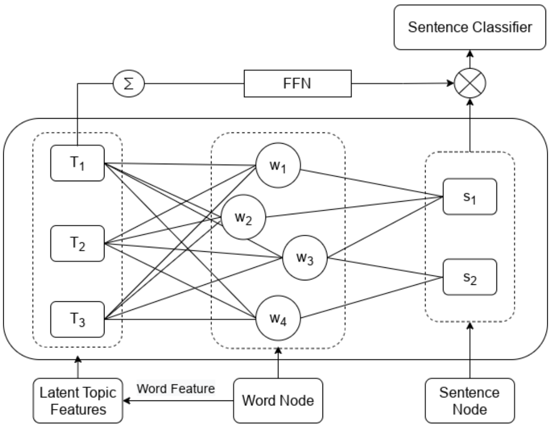


Figure 2: Overview of TopicHeterGraphSum for single document summarization. The initialized word node and sentence node features are processed following the work in (Wang et al., 2020). Furthermore, we provide latent topics as addition nodes into the Hetergraph.

Specifically, compared with previous works, the main idea of the proposed model is to enrich global information. Accordingly, instead of using neural models for generating latent topics, we first extract the initialized topic feature of each document using simple clustering methods (e.g., K-mean and GMM) of pre-trained word embeddings (Sia et al., 2020). In particular, the initialized topic feature is calculated as follows:

$$X_T = argmin \sum \begin{cases} \| c^{(i)} - x_j \|, Kmean \\ \theta_i f(x_j | c^{(i)}, \Sigma_i), GMM \end{cases} \tag{5}$$

where θ_i denotes topic proportions. $c^{(i)}$ and x_j represent the cluster center and word vector, respectively. Sequentially, the latent topics are put into a graph layer for extracting semantic information. Similar to sentence representation calculation in Eq. 3, the topic representation can be updated via modified GAT as follows:

$$\begin{aligned} U^1_{T\leftarrow w} &= GAT(H^0_T, H^0_w, H^0_w) \\ H^1_T &= FFN(U^1_{T\leftarrow w} + H^0_T) \end{aligned} \tag{6}$$

Each iteration contains word-to-sentence, sentence-to-word, and word-to-topic for the update process, which can be formulated as follows:

$$\begin{aligned} U^{t+1}_{w\leftarrow s} &= GAT(H^t_w, H^t_s, H^t_s) \\ U^{t+1}_{w\leftarrow T} &= GAT(H^t_w, H^t_T, H^t_T) \\ U^{t+1}_{w\leftarrow s,T} &= \sigma(U^{t+1}_{w\leftarrow s} + U^{t+1}_{w\leftarrow T}) \\ H^{t+1}_w &= FFN(U^{t+1}_{w\leftarrow s,T} + H^t_w) \\ U^{t+1}_{s\leftarrow w} &= GAT(H^t_s, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_s &= FFN(U^{t+1}_{s\leftarrow w} + H^t_s) \\ U^{t+1}_{T\leftarrow w}, A^{t+1}_{T\leftarrow w} &= GAT(H^t_T, H^{t+1}_w, H^{t+1}_w) \\ H^{t+1}_T &= FFN(U^{t+1}_{T\leftarrow w} + H^t_T) \end{aligned} \tag{7}$$

where $A_{T\leftarrow w}$ denotes the attention matrix from the word node to the topic node. Subsequently, the topic representation of the input document is calculated by combining all topic features, which are learned using GAT as follows:

$$\begin{aligned} \alpha_i &= \frac{\sum_{n=1}^{N_d} c(w_n) * A_{i,n}}{\sum_{j=1}^{K} \sum_{n=1}^{N_d} c(w_n) * A_{j,n}} \\ H_{T_d} &= \sum_{i=1}^{K} \alpha_i * H_{T_i} \end{aligned} \tag{8}$$

where $A_{i,j}$ indicates the amount of information word j contributes to topic i. $c(w_n)$ is the frequency of w_n

in the document, K is the number of topics and α_i refers to the level dominant of topic ith to the total document-topic. Sequentially, each sentence hidden state is integrated with the above topic vector to capture sentence-topic representation as follows:

$$H_{s_i,T_d} = FFN(H_{T_d}) \oplus H_{s_i} \tag{9}$$

Finally, the output sentence-topic representation is used for sentences classification by using cross-entropy loss as the training objective:

$$\mathcal{L} = \sum_{i=1}^{n} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{10}$$

3.2 Multi Documents Summarization

Currently, there are not many studies for multi-document summarization. The main challenge of MDS is that the input documents may differ in terms of main focus and point of view (Fabbri et al., 2019). Intuitively, enriching global information is able to improve the performance of the MDS problem in which latent topics, extracted from word nodes, are considered for whole sentences in the multi documents. Therefore, in this paper, we take MDS into account by extending our proposed HeterGNN model. Fig. 3 demonstrates the modification of our model for the multi documents. In particular, compared with the original model for SDS, there are several minor modifications. Firstly, latent topics are generated for covering the topics of whole relevant documents. In this regard, instead of combining all topic features for the topic representation, we keep each topic feature representation separately to maintain the information. Secondly, the word node and sentence node are generated by a set of relevant documents, which include a list of sentences and a set of unique words from multiple documents instead of a single document in the SDS problem. Specifically, supporting $D = \{d_1, d_2, ..., d_n\}$ denotes the set of each input multi documents, the output sentence-topic representation s_i is re-calculated as follows:

$$\begin{aligned} \bar{H}_{T_D} &= FFN(\Big\|_{k=1}^{K} H_{T_k}) \\ H_{s_i,T_D} &= \sigma(FFN(\bar{H}_{T_D} \oplus H_{s_i})) \end{aligned} \tag{11}$$

where K denotes the number of topics for the multiple documents and $\|$ represents the concatenation operation. Sequentially, the output matrix is transformed into the vector by a flattened layer for the final classification.

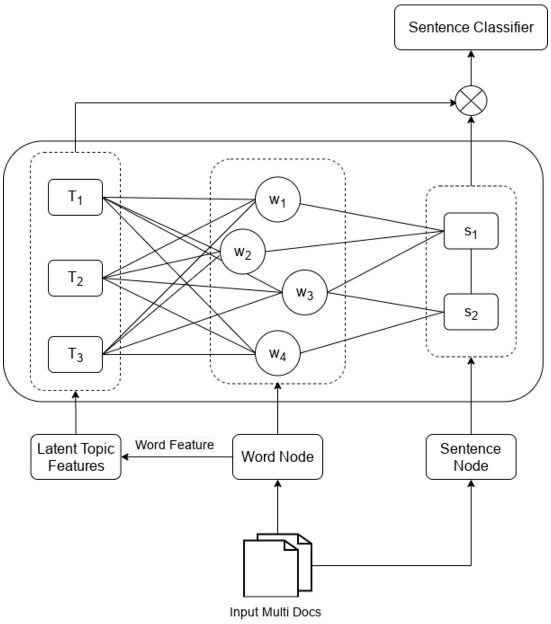


Figure 3: Overview of TopicHeterGraphSum for multi documents summarization.

4 Experiment

4.1 Experimental Setting

Datasets: Two benchmark datasets are considered for the evaluation such as CNN/DailyMail (Nallapati et al., 2016) (single document dataset) and Multi-News (Fabbri et al., 2019) (multi documents dataset). For the data processing, we use the same split as the work in (Wang et al., 2020). Specifically, the statistics of two benchmark datasets are illustrated in Tab. 1.

	CNN/Daily Mail	Multi-News
Train	287,227	44,972
Val	13,368	5,622
Test	11,490	5,622
Vocab Size	717,951	666,515

Table 1: Statistics of the evaluated datasets.

Hyperparameter Setting: Regarding the word node generation, the vocabulary is limited to 50,000. The tokens are initialized with 100 dimensions using Glove embeddings (Pennington et al., 2014). The multi-head of the GAT layer for word-to-sentence and word-to-topic is set to 4 and 1, respectively. The maximum number of sentences in each document is set to 100. The initialized dimensions of sentence embedding and topic embedding are set to 128 and 100, respectively. The dimension of the final output representation of all models is set to 64. Regarding the decoder process, we select top-3 for CNN/DailyMail and top-11 sentences for Multi-News following the performance of the validation set. Furthermore, n-gram Blocking (Liu and Lapata, 2019) is also taken into account to improve the performance. Specifically, we vary the values of n-gram from 3 to 6 in order to determine the best results. The number of latent topics is set to 5 both single and multi documents, respectively.

Baseline: For the SDS problem, several state-of-the-art non-pretrained models, which have recently introduced, are taken into account such as BANDDITSUM (Dong et al., 2018), JECS (Xu and Durrett, 2019), HER (Luo et al., 2019), Topic GraphSum (non-pretrained version) (Cui et al., 2020), HSG (Wang et al., 2020), and Multi GraS (Jing et al., 2021). Regarding the MDS problem, the most recent state-of-the-art methods using pre-trained models are proposed for abstractive summarization (Xiao et al., 2021). Consequently, we follow the reports in (Wang et al., 2020) to take the comparison. The proposed model, TopicHeterGraphSum (THGS) is executed with two versions, by adopting two clustering algorithms for initialized latent topic features, such as K-Mean (THGS-KMean) and GMM (THGS-GMM).

4.2 Main Results

Single Document Summurization: Table 2 shows the results of our evaluation on the CNN/DailyMail dataset. As result, our model outperforms the state-of-the-art models in this research field. Specifically, the results show that initialized features of latent topics by using GMM achieves better results than K-Mean.

Multi Document Summurization: Table 3 shows the results on the Multi-News dataset for the MDS problem. Specifically, the results indicate that enriching global information by using latent topics is able to improve the performance of the MDS problem.

Model	**R-1**	**R-2**	**R-L**
BANDITSUM	41.50	18.70	37.60
JECS	41.70	18.50	37.90
HER	42.30	18.90	37.90
Topic-GraphSum	41.93	19.15	38.22
HSG	42.95	19.76	39.23
Multi-GraS	43.16	20.14	39.49
THGS-Kmean (ours)	43.25	20.20	39.62
THGS-GMM (ours)	**43.28**	**20.31**	**39.67**

Table 2: Results on CNN/DailyMail dataset. Report results are obtained from respective papers. Bold texts indicate the best results in each column.

Model	**R-1**	**R-2**	**R-L**
TextRank	41.95	13.86	38.07
LexRank	41.77	13.81	37.87
PG-BRNN	45.27	15.32	41.38
Hi-MAP	45.21	16.29	41.39
HDSG	46.05	16.35	42.08
THGS-Kmean (ours)	46.60	16.81	42.63
THGS-GMM (ours)	**46.66**	**16.90**	**42.73**

Table 3: Results on Multi-News dataset. Reported results are obtained from (Wang et al., 2020). Bold texts indicate the best results in each column.

4.3 Results with Varying Hyperparameters

We execute experiments to evaluate the impact of important hyperparameters on the performance of the proposed model. Due to the limitation of our resources, we mainly focus on the Multi-News datasets.

Iteration Numbers: In order to select the best number of iterations of GAT, we measure the performance of different numbers of iterations on the validation sets. Table 4 shows the results in which the number of iterations ranges from 1 to 3. Accordingly, the larger number of iterations does not make a further substantial gain. Therefore, we select the number of iterations equal to 2 for the Multi-News

Iteration	**R-1**	**R-2**	**R-L**
1	46.16	16.57	42.05
2	46.66	16.90	42.73
3	46.53	16.87	42.67

Table 4: Results on Multi-News dataset with different number of iteration of GAT.

datasets. In the case of the CNN/DailyMail dataset, the number of iterations is set to 3, corresponding to the best performance.

Number of Sentences for Decode: Normally, the number of sentences for decoding is determined based on the average length of the human-written summaries. Accordingly, the average length of CNN/DailyMail and Multi-News are 3 and 9, respectively. However, we take this issue into account by varying the number of sentences. Table 5 shows the results of various numbers of sentences for decoding the Multi-News dataset. As result, we select the top-11 sentences for Multi-News datasets following the performance of the validation set.

Num. of Sent.	R-1	R-2	R-L
9	46.16	16.57	42.05
10	46.55	16.80	42.53
11	46.66	16.90	42.73
12	46.53	16.90	42.71
13	46.53	16.89	42.62

Table 5: Results on Multi-News dataset with different number of sentence for decoding.

N-gram Blocking: Trigram blocking is adopted to reduce redundancy for the decode process (Liu and Lapata, 2019). In this study, we vary the values of the n-gram from 3 to 6 to determine the best value. Accordingly, we set the value of n equal to 5 for the Multi-News dataset, which provides the best performance in terms of R-1 and R-L.

n-gram	R-1	R-2	R-L
3	46.04	16.07	42.04
4	46.61	16.73	42.65
5	46.66	16.90	42.73
6	46.60	16.93	42.68

Table 6: Results on Multi-News dataset with different numbers of n-gram blocking.

5 Conclusion and Future Work

We introduce a new method for the EDS problem by enriching global information using latent topics. Specifically, we first generate the latent topics using well-known clustering algorithms. The outputs are put into a HeterGNN as additional nodes for enriching the feature representations of sentences. The experiment on two benchmark datasets such as CNN/DailyMail and Multi-News of both SDS and MDS indicates the promising results of the proposed method in this research field. A major drawback of this study is that we use the same latent topic aggregation method for both SDS and MDS problems. Specifically, latent topics are suitable for the MDS problem, which has been proved in this study. However, since the complex relationship between word node and sentence node in multiple documents, a further investigation on exploiting the relationship between two types of nodes across multiple documents is able to improve the performance. Therefore, further exploitation of topic aggregation for the MDS problem is considered as our future work regarding this study.

References

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5360–5371. International Committee on Computational Linguistics.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* *Brussels, Belgium, October 31 - November 4, 2018*, pages 3739–3748. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3622–3631. Association for Computational Linguistics.

Baoyu Jing, Zeyu You, Tao Yang, Wei Fan, and Hanghang Tong. 2021. Multiplex graph neural network for extractive text summarization. *CoRR*, abs/2108.12870.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: human reading inspired extractive summarization. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3031–3041. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *CoRR*, abs/2109.10616.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1728–1736. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6209–6219. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. PRIMER: pyramid-based masked sentence pre-training for multi-document summarization. *CoRR*, abs/2110.08499.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong*

Kong, China, November 3-7, 2019, pages 3290–3301. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5021–5031. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir R. Radev. 2017. Graph-based neural multi-document summarization. In Roger Levy and Lucia Specia, editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 452–462. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Tweet Review Mining focusing on Celebrities by Machine Reading Comprehension based on BERT

Yuta Nozaki, Kotoe Sugawara, Yuki Zenimoto, Takehito Utsuro
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba,
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
{s2020776,s2120736,s2220753}_@_s.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

In this paper, we propose a method of mining tweets that represent reviews on celebrities by the machine reading comprehension model based on BERT. The purpose of this paper is to collect and aggregate reviews on tweets for celebrities in order to support the activities of fans of celebrities searching for information on celebrities' criticisms and impression trends on the Web. Specifically, we focus on the celebrities' names and adjective expressions that appear in tweets, and determine whether or not there is a relationship such as review and impression between the two words using the BERT machine reading comprehension framework. In the machine reading comprehension framework, the tweet is the context, the celebrity's name is the question, and an adjective that has a relationship such as review and impression with the celebrity is the answer. As the result of the evaluation experiment, the machine reading comprehension framework achieved fairly reliable performance.

1 Introduction

The purpose of this paper is to support the process by which TV drama viewers and celebrity fans search for information on critiques and interest trends in celebrities on the Web. In this paper, we propose a method for mining reviews of popular celebrities using tweets as the information source. Usually, when an event or an incident that is related to a popular celebrity occurs, a large number of tweets are posted, where, in those tweets, people express their own thoughts in the way they like. In Twitter, however, there exist no rule on the grammatical correctness of the posted tweets. Thus, this makes it unexpectedly difficult to correctly identify what people actually intend to express in their tweets, mainly due to the lack of grammatical correctness of tweet sentences (Sanguinetti et al., 2020). And, in the NLP community, in the case of other applications such as tweet sentiment analysis, it is quite common to avoid grammatically parsing tweet sentences, but to directly analyze sentiment of tweets (e.g, Nakov et al. (2016)).

Considering such a background, in this paper, we apply the BERT framework of machine reading comprehension (Figure 1) to tweet sentences, so as to detect the span of impression relations from text showing the review relations of celebrities. Specifically, first we collect tweets in which the name of the celebrity is included. We develop a dataset of tweets that are annotated with the span of adjectives indicating review relationships for the celebrity names co-occurring in the tweets. Next, in a machine reading comprehension framework, we train the BERT model (Devlin et al., 2019) to predict the spans that indicate reviews on previously unobserved celebrities.

Machine reading comprehension frameworks are often used in question answering systems. For example, Huang et al. (2020) proposed a machine reading comprehension model aimed at answering questions from Twitter, which is full of noisy, informal text. Also, Guo et al. (2021) proposed a model to detect spans mentioning health-related information from tweets. Xu et al. (2019) proposed a Review Reading Comprehension task that predicts spans from review text that are answers to user questions.

The model proposed in this paper, which is based on a machine reading comprehension framework, uses the keyword Q of the celebrity name as the question, and the tweet in which the keyword Q and the adjective A co-occur as the context C. If the review of the celebrity name Q is expressed by an adjective in the tweet, the corresponding adjective A is the output. If the adjective indicates an opinion on the celebrity Q, the corresponding adjective A is also the output. And, if the adjective does not indicate an opinion, "Not Review" is the output.

Specifically, first we collect tweets in which the name of the celebrity is included. We develop a dataset of tweets that are annotated with the span of adjectives indicating review relationships with the celebrity names co-occurring in the tweets. Next, in a machine reading comprehension framework, we train the BERT(Devlin et al., 2019) model to predict the spans that indicate reviews of unknown celebrities.

In the evaluation, first we develop the dataset as follows. In the dataset, for each tweet that mentions a specific celebrity, we annotate whether the tweet expresses an opinion on the celebrity by means of an adjective. First, we collect tweets that contain the name of a specific celebrity for a certain period of time. The tweets that contain adjectives with a certain frequency are randomly selected to be included in the dataset. Then, to each candidate tweet, it is assigned whether there is a review relationship between the celebrity name and the adjective, and the review mining dataset of 1,500 tweet instances is developed.

Using the developed dataset, we trained and evaluated the machine reading comprehension model (Figure 1) with BERT, where the results of comparison with the token classification model (Figure 2) with BERT as well as the classification model by SVM show that the machine reading comprehension model with BERT achieved the best performance. Especially, we further study to measure the performance of detecting spans that indicate the review relationship of previously unobserved celebrities. In this evaluation, it can be concluded that the machine reading comprehension model trained with tweets including previously observed celebrities' names is also effective in detecting the review relationship with previously unobserved celebrity names. This is quite contrastive in that the token classification model is not capable of detecting the review relationship with previously unobserved celebrity names.

We also investigate the effect of the number of adjectives co-occurring within a tweet on the performance of the machine reading comprehension model. From this analysis, it is shown that, in the case where the number of co-occurring adjectives is one, the performance of review relation detection is much higher than the case where the number of co-occurring adjectives is two or more. However, even in the case where the number of co-occurring adjectives is two or more, its precision is over 0.5, which is more or less satisfactory performance and this result confirms the effectiveness of the proposed approach.

2 Developing the Review Mining Dataset

2.1 Collecting Tweets

In this paper, we collected tweets on 5 celebrities[1] who are very popular in Japan. We collected tweets on each celebrity to develop the dataset of candidate tweets. First, we use the Twitter Search API[2] to collect tweets that contain each celebrity's name as the keyword, where the numbers of collected tweets are shown in Table 1[3]. Then, with the results of morphological analysis by JUMAN++[4] on the collected tweets, we transform each adjective into its representative notation (base form) and obtain the frequency statistics of the adjectives that co-occur with each celebrity name. Next, for each celebrity name, the most frequent 30 adjectives are used to collect candidate tweets for developing the dataset. Specifically, 10 tweets were randomly selected for each

[1] Satomi Ishihara, Haruma Miura, Masato Sakai, Yuko Takeuchi, and Ryoko Yonekura,

[2] https://developer.twitter.com/en/docs/ tweets/search/api-reference/get-search-tweets

[3] For roughly around the period from July to October, 2020, the numbers of collected tweets (not including retweets) are 555,701 for Satomi Ishihara, 130,734 for Haruma Miura, 61,108 for Masato Sakai, 33,7672 for Yuko Takeuchi, and 14,869 for Ryoko Yonekura.

[4] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++

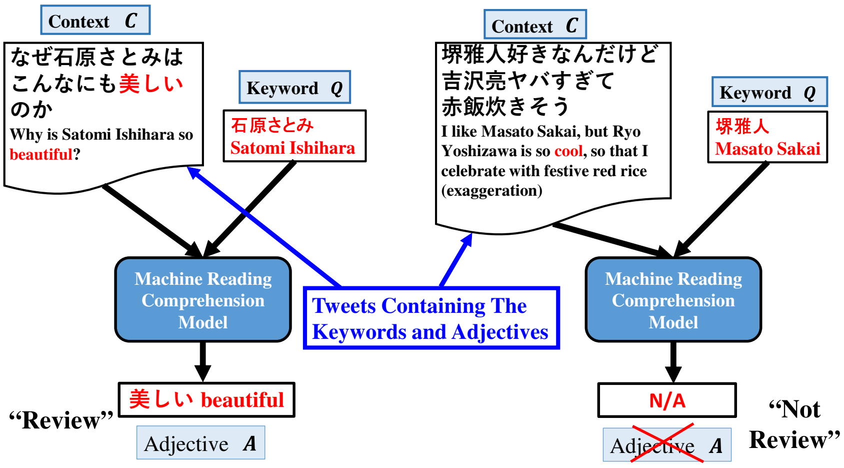

Context C
なぜ石原さとみはこんなにも美しいのか
Why is Satomi Ishihara so beautiful?
Keyword Q
石原さとみ
Satomi Ishihara
Machine Reading Comprehension Model
Tweets Containing The Keywords and Adjectives
美しい beautiful
“Review”
Adjective A
Context C
堺雅人好きなんだけど吉沢亮ヤバすぎて赤飯炊きそう
I like Masato Sakai, but Ryo Yoshizawa is so cool, so that I celebrate with festive red rice (exaggeration)
Keyword Q
堺雅人
Masato Sakai
Machine Reading Comprehension Model
N/A
Adjective A
“Not Review”

Figure 1: Framework of Machine Reading Comprehension Model for Mining Reviews on Celebrities represented by Adjectives

adjective. Then, 1,500 tweets in total are used as candidate tweets for developing the dataset.

2.2 Distribution of Adjectives per Celebrity

For each of the three celebrity names “Satomi Ishihara”, “Masato Sakai”, and “Ryoko Yonekura”, Figure 3 shows the statistics of the frequencies of the most frequent 30 adjectives. More or less general adjectives that are common among celebrity names such as “good” and “amazing” are observed in the highest ranks for “Satomi Ishihara” and “Masato Sakai”. Other adjectives, however, are mostly specific to each celebrity name. Those celebrity specific ones sometimes represent events closely related to each celebrity such as the marriage of “Satomi Ishihara”. Those celebrity specific ones include “envy”, “congratulation”, and “happy” for “Satomi Ishihara”, “reliable”, “painful”, and “unique” for “Masato Sakai”, “noisy”, “close” and “annoying” for “Ryoko Yonekura”[5].

2.3 Criteria on Annotating Reviews on Celebrities represented by Adjectives

For the selected candidate tweets, we develop the review mining dataset by annotating whether or not the tweets show review relations between the celebrity name and adjectives. The results of annotating 1,500 candidate tweets based on the criteria can be divided into the following three classes.

- ”Review” · · · For the keyword celebrity name, only one adjective in the tweet indicates the review relationship with the keyword celebrity name. Here, the adjective does not have to be the one specified at the time of tweet collection.
- ”Not Review” · · · With the keyword celebrity name, none of the adjectives in the tweet show a review relationship[6].
- Others · · · With the keyword celebrity name, two or more adjectives in the tweet show a review relationship.

Among the 1,500 tweets[7], the tweets corresponding to others are excluded, and the same number of

[5] In the evaluation, we apply the pre-trained BERT machine reading comprehension model, which helps to avoid overfit to adjectives seen in fine-tuning. Actually, we confirmed in the supplementary evaluation we omit due to the space restriction in this paper, that the model is capable of answering adjectives unseen in fine-tuning in the machine reading comprehension framework of this paper.

[6] Examples of “Not Review” include tweets that indicate a review relationship with a celebrity other than the keyword celebrity or a TV drama in which the celebrity appears, and tweets that indicate a review relationship with a part of speech other than adjectives.

[7] As a result, as shown in Table 1, the number of tweets in ”Review” totaled 636 of which 112 for Satomi Ishihara, 147 for Haruma Miura, 131 for Masato Sakai, 126 for Yuko Takeuchi, 120 for Ryoko Yonekura. The number of tweets in ”Not Review” totaled 864 of which 188 for Satomi Ishihara, 153 for

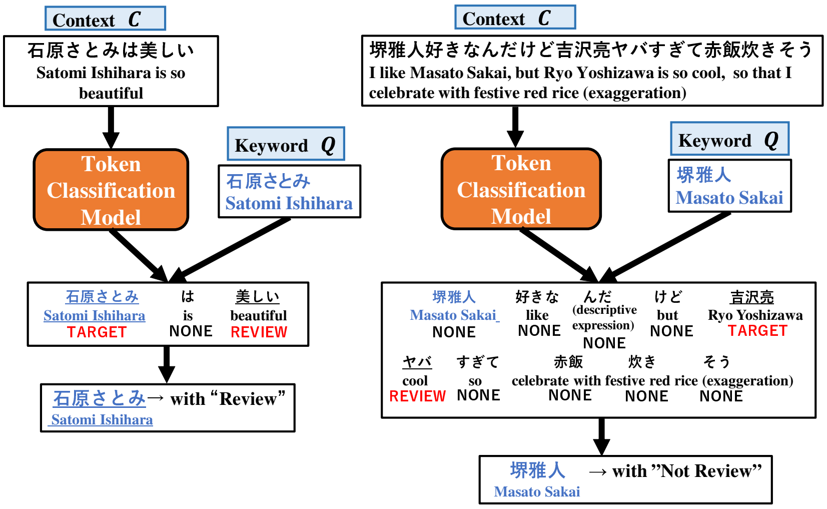


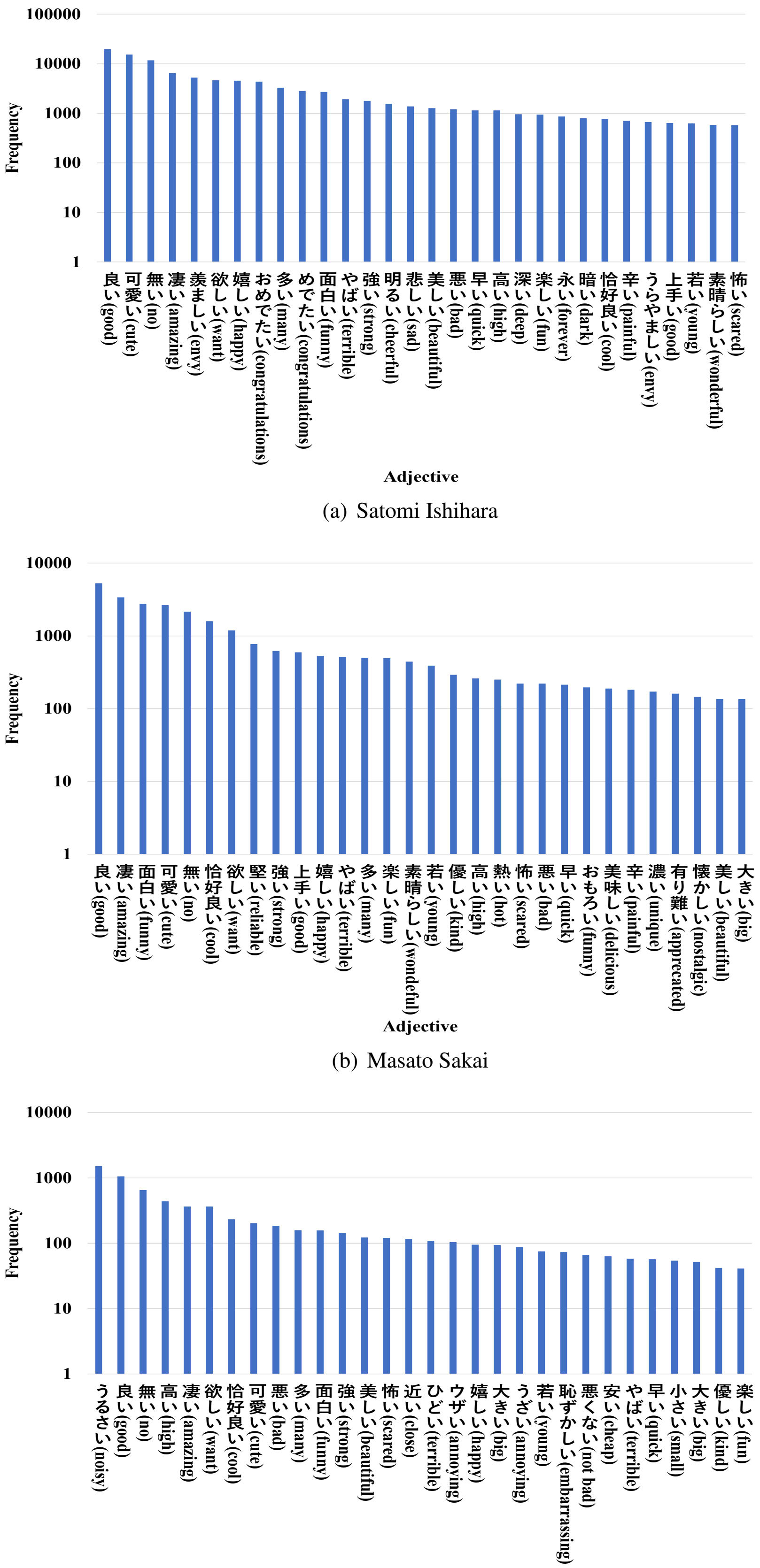
Figure 2: Framework of Token Classification Model for Mining Reviews on Celebrities represented by Adjectives

Table 1: Numbers of Collected Tweets and those for Evaluation

celebrity name	number of tweets	Number of tweets excluding retweets	collection period	Number of training and evaluation tweets		
				Review for 1 adjective only.	not Review	total
Satomi Ishihara	1,677,692	555,701	2020/8/2 ～ 2020/10/24	112	188	300
Haruma Miura	557,796	130,734	2020/7/22 ～ 2020/8/15	147	153	300
Masato Sakai	161,404	61,108	2020/8/2 ～ 2020/10/24	131	169	300
Yuko Takeuchi	868,979	337,672	2020/9/27 ～ 2020/10/24	126	174	300
Ryoko Yonekura	37,445	14,869	2020/8/2 ～ 2020/10/24	120	180	300
total	-	-	-	636	864	1,500

Table 2: Number of Adjectives Co-occurring in a Tweet (%)

Review or Not Review	Number of co-occurring adjectives in a tweet		total
	1	≥ 2	
Review for 1 adjective only.	335 (22.3)	300 (20.0)	635 (42.3)
Not Review	487 (32.5)	378 (25.2)	865 (57.7)
total	822 (54.8)	678 (45.2)	1,500 (100)



(a) Satomi Ishihara

(b) Masato Sakai

(c) Ryoko Yonekura

Figure 3: Statistics of the Frequencies of Adjectives Co-occurring with Celebrity Names (30 Types of Adjectives with the Highest Frequencies)

tweets are additionally collected from the tweets in which the pair of celebrity name c and adjective a co-occur.

3 Review Mining Models

3.1 Machine Reading Comprehension Model

For modeling with the machine reading comprehension model, we use the question-answering framework of Figure 1. In this framework, tweets which contain a celebrity name are used as the context C, the celebrity name is used as the keyword Q, and the adjective A is the output if the adjective indicates a review relation to the celebrity name Q in the tweet. If all the adjectives in the tweet do not indicate a review relationship with the celebrity name Q, nothing is output. With this framework, it is judged whether or not there exist any review relation with the celebrity name Q in the tweet.

Specifically, in the case of the tweet in the ”Review” class, we store the example as a tuple of the context C, the question as the keyword Q for the celebrity name, and the token position L_A of the answer adjective A indicating the review relationship with Q.

Context C**:** the tweet in which the keyword Q is co-occurring with the adjective A, which has the review relationship with Q.

Question Q**:** the keyword Q as the celebrity name

Token position L_A **of answer** A **:** the token position of the adjective A, which has the review relation with Q.

In the case of the tweet in the “Not Review” class, on the other hand, we store the example as a tuple of the context C for the machine reading comprehension model (Devlin et al., 2019) training, the question Q as the keyword for the celebrity name, and answer A' = N/A.

Context C**:** the tweet where the keyword Q and all adjectives in the tweet do not have review relationship.

Question Q**:** the keyword Q as the celebrity name

Answer A'**:** N/A

3.2 Token Classification Model

Weinzierl and Harabagiu (2020) divided tweets into tokens and predicted what events each token was related to using a multi-class classification model. In this paper, we also divide tweets into tokens and build a token classification model that predicts whether each token indicates the name of a celebrity, a review of that celebrity, or something else.

For modeling with the token classification model, we use the named entity recognition framework of BERT (Devlin et al., 2019). The token classification model in this paper is shown in Figure 2. In the training data of the token classification model, the sequence of all morphemes in a tweet which contains a celebrity name is denoted as $L_1,\ldots,L_n$, and each morpheme L_i ($i = 1, \ldots, n$) is regarded as a token, and is annotated with one of the 3 classes “REVIEW”, “TARGET” and “NONE”. The “REVIEW” class is a class that indicates an opinion on the celebrity, the “TARGET” class is a class that indicates the name of the celebrity, to whom the opinion of the “REVIEW” class is directed, and the “NONE” class is a class that indicates other morphemes.

As shown in Figure 2, when testing, given the tweet which is denoted as the context C (containing the query keyword Q of the celebrity name) and the query keyword Q of the celebrity name, the token classification model predicts the class of each morpheme L_i ($i = 1, \ldots, n$) into 3 classes “REVIEW”, “TARGET” and “NONE”. Then, if the morpheme L_j ($1 \leq j \leq n$) of the query keyword Q of the celebrity name is predicted as “TARGET”, and there exists at least one morpheme L_k ($1 \leq k \leq n$) that is predicted as “REVIEW”, then, the overall output as “with REVIEW” is predicted.

4 Evaluation

4.1 The Procedure

In this paper, we used PyTorch implementation of BERT (Devlin et al., 2019) for both the machine reading comprehension model and the token classification model. For BERT, we used the NICT BERT Japanese Pre-trained model[8], which was pre-trained using the entire Japanese Wikipedia except for the

Haruma Miura, 169 for Masato Sakai, 174 for Yuko Takeuchi, 180 for Ryoko Yonekura. The number of tweets excluded as others totaled 169 of which 20 for Satomi Ishihara, 64 for Haruma Miura, for 21 Masato Sakai, for 44 Yuko Takeuchi, for 20 Ryoko Yonekura.

[8] https://alaginrc.nict.go.jp/nict-bert/index.html

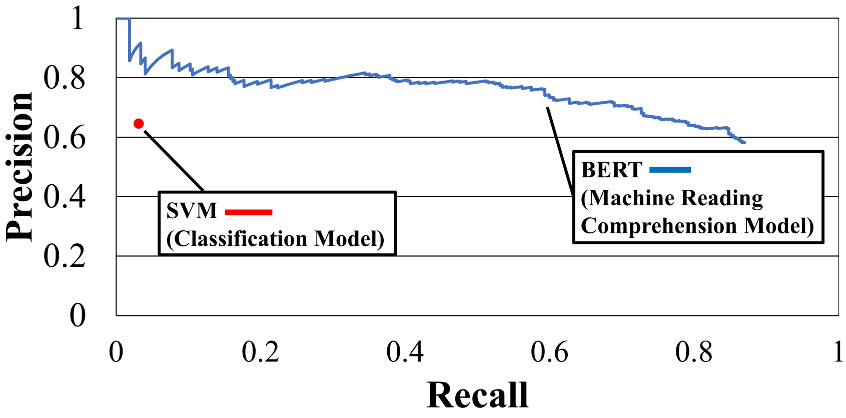


Figure 4: Evaluation Results where the Model Output is “Review” (5-fold Celebrities Cross-Validation)

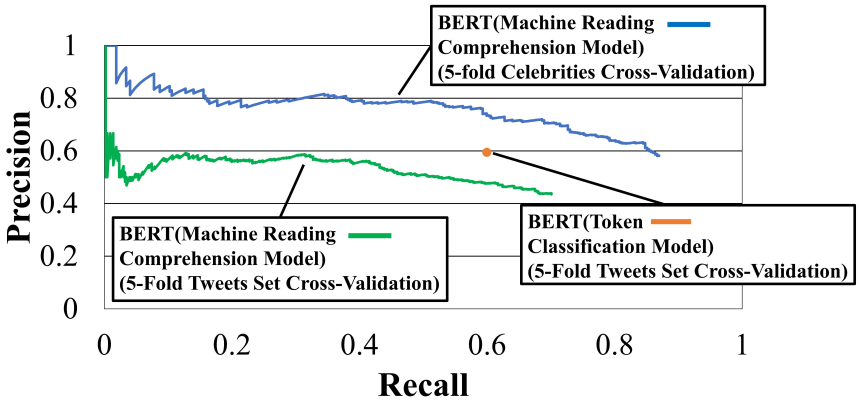


Figure 5: Comparative Evaluation Results of 5-fold Celebrities Cross-Validation and 5-fold Tweets Set Cross-Validation (where the Model Output is “Review”)

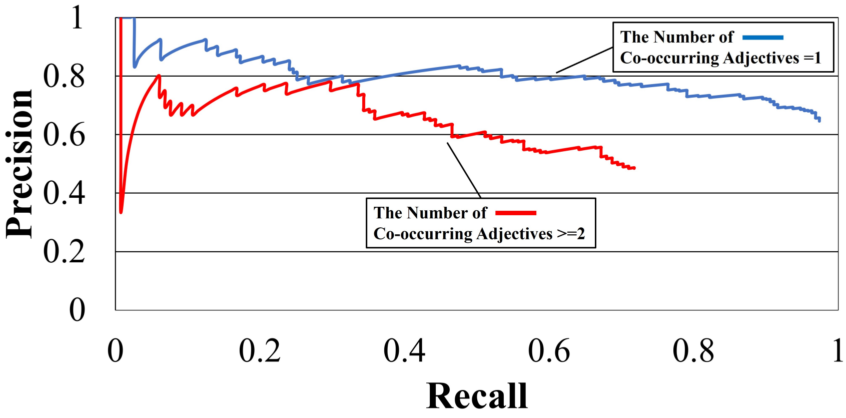


Figure 6: Evaluation Results per the Numbers of Co-occurring Adjectives in a Tweet (where the Model Output is “Review” and the Model is BERT (Machine Reading Comprehension Model))

title. The context C is segmented into morpheme sequences by JUMAN++. Then, based on the BERT specification, the WordPiece module[9] was applied to segment the context C further into subword units with a vocabulary of 32,000. For fine-tuning of the machine reading comprehension and the token classification models, we used the modules[10] of Huggingface[11]. For the evaluation, we conducted two types of cross-validation, namely, 5-fold celebrities cross-validation and 5-fold tweets set cross-validation. In 5-fold celebrities cross-validation, we trained on the data of 4 of the 5 celebrities and evaluated on the data of the remaining one celebrity. The 5-fold tweets set cross-validation is the usual 5-fold cross-validation for the whole set of collected tweets. The whole set of collected tweets was randomly divided into 5 subsets regardless of the name of the celebrity, and the training was conducted on 4 of the 5 subsets and the evaluation was conducted on the remaining one subset.

As a baseline, we modeled the text classification model[12] with SVM. The implementation was done using `scikit-learn`. As the feature scaling of the dataset, each tweet was vectorized using tf-idf to train a linear SVM classification model[13], and 5 celebrities cross-validation was performed.

4.2 Evaluation Result

With the "Review" class as the positive examples, Figure 4 plots the recall-precision curve of the machine reading comprehension model as "BERT (Machine Reading Comprehension Model)", where the lower bound p_0 of the output probability of BERT softmax function is changed in descending order[14]. Since the output probability of BERT softmax function is reliable as the confidence of BERT softmax function, the precision decreases as the recall increases. Figure 4 also plots precision and recall for the baseline as "SVM (Classification Model)"[15]. The figure shows the results of the 5-fold celebrities cross-validation. Here, in the 5-fold celebrities cross-validation, the token classification model cannot predict the query celebrity name that is not observed in training as "TARGET". Thus, we omit the plot for the token classification model. It is obviously shown that the machine reading comprehension model outperforms the baseline.

For the machine reading comprehension model, Figure 5 compares the 5-fold celebrities cross-validation with 5-fold tweets set cross-validation. Figure 5 also plots precision and recall for the token classification model as "BERT (Token Classification Model)". For the the machine reading comprehension model, the 5-fold celebrities cross-validation outperforms the 5-fold tweets set cross-validation, suggesting that the machine reading comprehension model is a model that does not over-fit the features of each celebrity. From this result, it can be concluded that the machine reading comprehension model trained with tweets including previously observed celebrities' names is also effective in detecting the review relationship with previously unobserved celebrity names. For the case of 5-fold tweets set cross-validation, the precision of the token classification model is relatively low compared with the machine reading comprehension model in the high confidence range of prediction.

Table 2 shows the statistics of the number of adjectives co-occurring in a tweet. This table shows that there exist certain percentages of the cases where the number of co-occurring adjectives is two or more, while one of those co-occurring adjectives has the review relationship with a celebrity name. Figure 6 further compares the recall-precision curves of the following two cases: (i) where the number of co-occurring adjectives is one, and (ii) where the number of co-occurring adjectives is two or more. This result shows that, in the case where the number of co-occurring adjectives is one, the performance of review relation detection is much higher than the case where the number of co-occurring ad-

[9] `tokenization.py`

[10] `run_squad.py` and `run_ner.py` were used respectively where the number of epochs as 2, the batch size as 8, and the learning rate as 0.00003.

[11] transformers-2.2.1

[12] The model takes "the celebrity name token sequence + [SEP] token + the tweet token sequence" as the input and classifies the input into 2 classes "Review" or "Not Review".

[13] C=1.0

[14] In this case, the recall does not reach 1 because there exist cases where the answer span predicted by the model does not match the reference answer. Thus, we only show the recall-precision curves, whereas we do not show ROC-curves in this paper.

[15] Although it is also possible to plot the recall-precision curve for SVM classification model, we omit it since the recall of SVM classification model is too low.

jectives is two or more. However, even in the case where the number of co-occurring adjectives is two or more, its precision is over 0.5, which is more or less satisfactory performance and this result confirms the effectiveness of the proposed approach.

5 Related Work

Compared with our analysis of Twitter mentions of celebrities, as a related work, Wiegmann et al. (2019) studied author profiling of celebrities in Twitter. Wiegmann et al. (2019) collected Wikipedia entries and Twitter feeds for 71,706 celebrities and developed a corpus containing an average of 29,968 words and a maximum of 239 personal traits per celebrity. They developed a model to predict gender and occupation from tweets using deep learning methods.

Span prediction has been also studied in previous tasks. For example, Alhuzali and Ananiadou (2021) proposed a model that casts the emotion classification task as span prediction. In the context of named entity recognition, Fu et al. (2021) studied the strengths and weaknesses of the span prediction model and compared it with the sequence labeling framework. In this paper, we employed the machine reading comprehension model based on span prediction to extract adjectives that indicate a review relationship with a celebrity name.

6 Conclusion

In this paper, we focus on the celebrities' names and adjective expressions that appear in tweets, and determine whether or not there is a relationship such as review and impression between the two words using the BERT machine reading comprehension framework. We also compared 5 celebrities cross-validation with 5-fold cross-validation, which showed that the machine reading comprehension model is robust enough not to overfit the features of each celebrity.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21H00901.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online, April. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online, August. Association for Computational Linguistics.

Yuting Guo, Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2021. Pre-trained transformer-based classification and span detection models for social media health applications. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 52–57, Mexico City, Mexico, June. Association for Computational Linguistics.

Rongtao Huang, Bowei Zou, Yu Hong, Wei Zhang, AiTi Aw, and Guodong Zhou. 2020. NUT-RC: Noisy user-generated text-oriented reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2687–2698, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France, May. European Language Resources Association.

Maxwell Weinzierl and Sanda Harabagiu. 2020. HLTRI at W-NUT 2020 shared task-3: COVID-19 event extraction from Twitter using multi-task hopfield pooling. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 530–538, Online, November. Association for Computational Linguistics.

Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618, Florence, Italy, July. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2324–2335. Association for Computational Linguistics.

Aspect-based Sentiment Analysis for Vietnamese Reviews about Beauty Product on E-commerce Websites

Quang-Linh Tran, Phan Thanh Dat Le, Trong-Hop Do
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{18520997, 18520570}@gm.uit.edu.vn, hopdt@uit.edu.vn

Abstract

Millions of reviews are generated on e-commerce platforms and analyzing them deeply brings a lot of useful information for sellers and buyers. This paper deals with the aspect-based sentiment analysis problem to analyze the aspect and polarity of Vietnamese reviews about beauty products on e-commerce websites. The contribution of this paper is three-fold. Firstly, we introduce a dataset containing 16,227 reviews about lipsticks. There are 32,775 pairs of aspects and sentiments in this dataset. Besides, we conduct some baseline experiments using Deep Learning-based models to build detection and classification models for extracting the aspects of reviews and classifying the sentiment of each aspect. In addition, a comprehensive comparison is also performed to see whether single-task learning or multi-task learning is the better approach to predict the aspect and sentiment of reviews. The experimental results show that the BiGRU+Conv1D model in the single-task learning approach outperforms others with the F1-score in the aspect detection task at 98.09% and 91.01% for the sentiment classification task.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a task in Natural Language Processing (NLP). Instead of only extracting the polarity of the text in sentiment analysis, ABSA analyzes the sentiment more deeply by showing the polarity of each aspect of the text. In reviews on e-commerce platforms, a review contains more than one aspect so ABSA is more robust to analyze the polarity than sentiment analysis. For example, in lipstick reviews, the customers do not always give general reviews about the products, they tend to review based on the aspect of the lipstick such as price, colour, et cetera. The customers may like the colour, but the texture of the lipstick is so bad that they are disappointed with this. This is a reason why ABSA is really necessary to understand more deeply what customers think about the products.

With the explosion of e-commerce, millions of reviews are generated every day from customers. They leave a lot of useful information for not only the sellers to understand what their customers like and dislike about their products but also for other customers to read and consider previous buyers' experiences before making a purchase. There are many kinds of products in e-commerce, but beauty products are one of the most favorable products to buy online. In addition, the reviews of beauty products, especially about the lipsticks contain a lot of aspects, from price, staying power to colour, thanks to which, several analyses can be conducted to deeply understand the attitude of customers toward the lipstick. This is the reason why we choose reviews about lipsticks to build a dataset to deal with ABSA problem in e-commerce reviews.

There are three main contributions of this paper. Firstly, a novel Vietnamese dataset about reviews of lipsticks in e-commerce platforms for the aspect-based sentiment analysis task is created to handle the problem. To the best of our knowledge, there is no dataset for reviews about beauty products in Vietnamese and the size of this dataset is also bigger than some other Vietnamese datasets for ABSA. Secondly, we propose an effective deep learning model architecture to detect the aspect and corresponding sentiment of reviews. Finally, a comprehensive comparison between single-task learning and multi-task learning is conducted to find the best approach for the aspect-based sentiment analysis problem.

2 Related Works

Aspect-based sentiment analysis has drawn a lot of attention in recent years. Several workshops such as SemEval2014 (Pontiki et al., 2014), SemEval2015 (Pontiki, Galanis, Papageorgiou, Manandhar, & Androutsopoulos, 2015), SemEval2016 (Pontiki et al., 2016), VLSP2018 (H. T. Nguyen et al., 2018) were organized to find the best solution for aspect-based sentiment analysis problem. There are many approaches to resolving this problem. Single task ABSA contains several sub-tasks such as aspect category detection, and aspect sentiment classification (Zhang, Li, Deng, Bing, & Lam, 2022). (Zhou, Wan, & Xiao, 2015) achieved the F1-score of 90.10% in aspect category detection in the SemEval2014 dataset (Pontiki et al., 2014) with representation learning. (Wang, Huang, Zhu, & Zhao, 2016) used attention-based LSTM for aspect sentiment classification. Besides single task ABSA, compound ABSA is also an effective approach. In this approach, aspect category sentiment analysis is the task to extract aspects and the corresponding sentiment simultaneously. (He, Lee, Ng, & Dahlmeier, 2019) proposed an interactive multi-task learning network for extracting aspects and sentiment of documents.

In Vietnamese, there are several datasets about sentiment analysis in many domains. (H. T. Nguyen et al., 2018) published the dataset SA-VLSP2018 for aspect-based sentiment analysis about hotel and restaurant domains in the VLSP workshop. In addition, (K. T.-T. Nguyen et al., 2021) applied span detection for aspect-based sentiment analysis and get the performance at 62.76% F1-score for the dataset UIT-ViSD4SA. (K. V. Nguyen, Nguyen, Nguyen, Truong, & Nguyen, 2018) published the dataset UIT-VSFC, which consists of over 16,000 sentences of feedback from students. This dataset is used for sentiment classification and topic classification.

3 Dataset

3.1 Task Definition

We built a Vietnamese dataset for the aspect-based sentiment analysis task. This dataset contains 16,227 Vietnamese reviews about 9 lipsticks from Shopee[1]. There are 2 sub-tasks in this dataset: aspect detection and sentiment classification. In the aspect detection sub-task, we focus on finding the aspects mentioned in the review. There are 7 aspects: SMELL, PRICE, SHIPPING, COLOUR, PACKING, TEXTURE, STAYING POWER and 1 aspect indicating spam: OTHERS. Table 1 shows the definition of all aspects. Another sub-task is classifying the polarity of these aspects into: Positive, Neutral or Negative. We split the dataset into three sets: 12,981 reviews for training, 1,623 reviews for validation, and 1,623 reviews for testing. The training and validation will help us to build detection and classification models, and the test set is used to evaluate the performance of the models.

3.2 Annotation process

We split our annotation process into 2 main phases, including training phase and labelling phase. There are 5 sub-phases in the training phase, each sub-phase has 200 reviews. We use Cohen's Kappa (Cohen, 1960) score to measure inter-annotator agreement as the metric for calculating the quality of annotation. When the score between annotators in each sub-phase is higher than 0.65, we stop training our annotators and move to the next sub-phase. Figure 1 illustrates the Cohen's Kappa score between 6 annotators of 5 training sub-phases.

After the training phase is done, annotators are able to label the rest of the dataset in the second phase, the labeling phase. There are 2 sub-phases, each sub-phase has about 7,500 reviews. Three annotators are responsible for annotating 7,500 separately. For disagreed reviews, we automatically choose the label which is chosen by 2/3 annotators.

3.3 Statistics

Our dataset contains 16,227 reviews, including 7 sentiment aspects and 1 aspect indicating spam reviews. Table 2 shows some examples of reviews in

[1] https://shopee.vn/

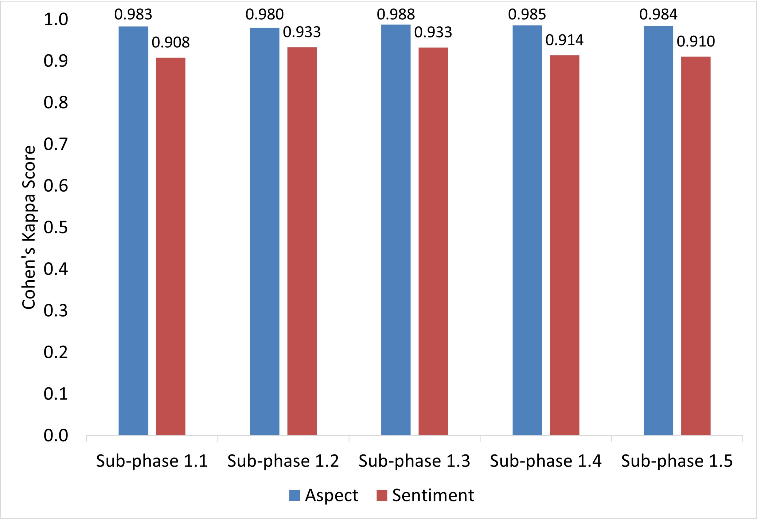


Figure 1: Inter Annotation Agreement Score of the training phase.

Aspect	Definition
SMELL	Reviews mention the scent of the lipstick.
COLOUR	Reviews mention lipstick color such as dark or light
TEXTURE	Reviews mention the characteristic and quality of the lipstick such as information about the moisture and dryness of the lipstick.
PRICE	Reviews mention the price of lipsticks, whether it is affordable or not.
STAYINGPOWER	Reviews mention the adhesion of lipsticks on the lip.
SHIPPING	Reviews mention the delivery service such as the time, the shipper's attitude.
PACKING	Reviews mention the quality of packing, whether the lipsticks are well packed or not.
OTHERS	Spam reviews.

Table 1: Aspect Definition.

our dataset and the corresponding aspects and polarities. Figure 2 illustrates the distribution for each aspect and sentiment in the dataset. Overall, the sentiment Positive accounts for the largest part in all aspects, except the aspect OTHERS. In addition, the number of reviews having aspect COLOUR is over 7000 reviews, accounting for nearly 50% of reviews, which shows the high concern of customers about this aspect when buying lipsticks. The high imbalance between sentiments can be seen in the aspect PRICE and PACKING. The reason is the price on e-commerce platforms is really competitive and cheaper than in physical stores, so this tends to receive good reviews. Meanwhile, the PACKING aspect receives good reviews because this is an aspect that sellers can control and manage to receive a good initial impression from buyers. Table 3 shows overview statistics of our dataset.

4 Aspect-based Sentiment Analysis model

This section gives information about the model that is used for the aspect-based sentiment analysis problem. The model has three main components: input layer & hidden layers, output layer for single-task learning, and output layer for multi-task learning approach. For the input layer and hidden layers in subsection 4.1, the model architecture is the same between single-task learning and multi-task learning. Based on the learning approach, the output layer can be different. The subsection 4.2 and 4.3 give more details about the output layers of two approaches.

Review	Aspect	Polarity
về mẫu mã khá cute mới đánh thử thì thấy oke màu lên môi đẹp mùi cũng thơm *(the sample is so cute, I just used it, the colour* *is great on my lip and the smell is also fragant too.)*	SMELL, COLOUR	SMELL:Positive, COLOUR:Positive
Công dụng: màu lì Son đẹp lắm ạ đóng gói rất kĩ hàng không bị móp méo xài rất mướt *(Uses: the colour is very adhesive, the beautiful* *lipsticks. Very well packed, no dents)*	STAYINGPOWER	STAYINGPOWER: Positive
Giao hàng lâu. Chờ mòn mỏi luôn í. Chak do hàng quốc tế. Màu khá ok. Sẽ ủng hộ shop *(Long delivery. Tired of waiting. Maybe because of* *international delivery. The colour is ok.* *I will support the shop.)*	COLOUR, SHIPPING	COLOUR:Positive, SHIPPING:Negative

Table 2: Several examples of the dataset.

Set	Review	Avg aspect/ review	Positive	Neutral	Negative	Total sentiment
Train	12,981	2.02	18,694	3,822	3,715	26,231
Dev	1,623	2.01	2,336	462	466	3,264
Test	1,623	2.02	2,298	511	471	3280

Table 3: Statistics about the experimental dataset.

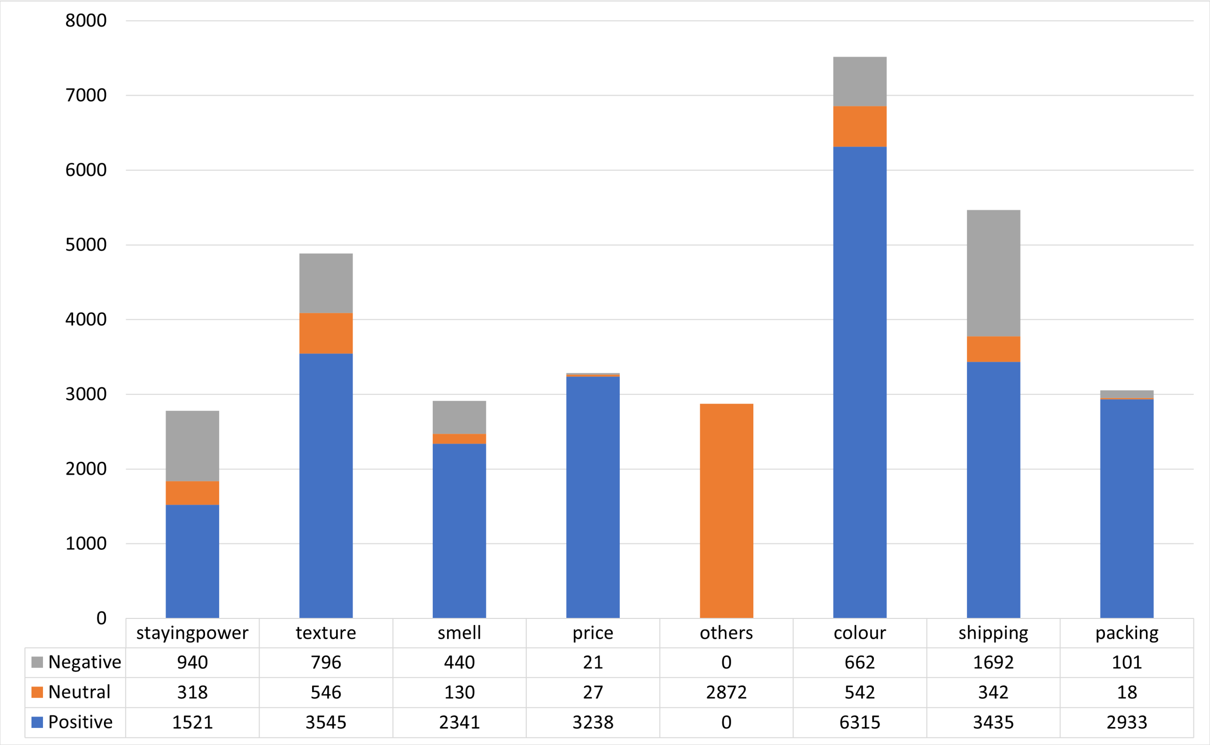
4.1 Input and Hidden layers

The input and hidden layers are illustrated in the figure 3. After pre-processing the reviews, a tokenizer layer will be used to convert from words to indexes based on the vocabulary index. The ELMO pre-trained word embedding (Vu, Vu, Tran, & Jiang, 2019) is used as the initialization for the embedding layer and this layer creates a representative vector for every word. The SpatialDropout1D layer helps to reduce the overfitting problem. A Bi-LSTM (Hochreiter & Schmidhuber, 1997) and Bi-GRU (Chung, Gulcehre, Cho, & Bengio, 2014) are used parallelly to obtain as much information as possible. The Bi-LSTM layer can save valuable information from the beginning of the reviews and utilize it to predict the label. After the Bi-LSTM or Bi-GRU layers, the Conv1D layers convert multi-dimensional matrices to 1D matrices and GlobalMaxPooling and GlobalAveragePooling will extract the maximum element of the matrices as well as the average element. The reason why parallel RNN-based neuron networks are used is that they can extract more features than a single neuron network. All pooling layers are concatenated and go through a dense layer before passing to output layers. It is worth noting that depending on the type of learning, which will be described at 4.2 and 4.3 below, the output layer can be different.

To find the best model architecture, we stack layers one by one from the Bi-LSTM layer or Bi-GRU layer to Bi-LSTM+Conv1D or Bi-GRU+Conv1D to the full layer version as in figure 3.

4.2 Output layer for Single-Task Learning approach

Single-Task Learning (STL) is a type of Deep Learning, in which a model is only specific to a task. In ABSA, there are many sub-tasks and they can be categorized as single-task learning. There are two main sub-tasks in ABSA, which are aspect category detection, and aspect sentiment classification (Zhang et al., 2022). After the aspect category detection model extracts the aspects in a review, the sentiment estimation model will predict the polarity of that aspect in the review. Because of this, for every aspect, a sentiment classification model needs to be built to estimate the polarity of that aspect.



	stayingpower	texture	smell	price	others	colour	shipping	packing
■ Negative	940	796	440	21	0	662	1692	101
■ Neutral	318	546	130	27	2872	542	342	18
■ Positive	1521	3545	2341	3238	0	6315	3435	2933

Figure 2: Aspect and Sentiment distribution in the dataset.

For our problem, there is one multi-label classification model to detect whether one or more aspects exist in the reviews. After that, seven sentiment classification models for seven aspects will be used to predict the sentiment of that aspect in the review. However, if the aspect detection model does not detect an aspect, the corresponding sentiment classification model will not estimate the polarity of that aspect in the reviews.

The output for the single-task learning approach has 8 units corresponding to 8 aspects for the aspect detection model or 3 units corresponding to 3 polarities for the sentiment classification models. The figure 4 illustrates the example of the sentiment classification for the aspect COLOUR. We have to build 6 other models like this for the sentiment classification and 1 model for the aspect detection task.

4.3 Output layer for Multi-Task Learning approach

In the multi-task learning (MTL) approach, the aspects and corresponding sentiment are predicted simultaneously. This approach has been used in a lot of previous research (He et al., 2019), (Luo, Li, Liu, & Zhang, 2019) and it proves its effectiveness in the aspect-based sentiment analysis task. Inspirited by these researches, we design a multi-task recurrent neural network for our own problem. The input layer and hidden layers in the multi-task model are similar to model architecture in 4.1, and they serve as the shared layers, however, for the task-specific layers, there are an aspect detection task and seven sentiment classification tasks. An example of a multi-task learning model is illustrated in the figure 5. The aspects and corresponding sentiment will be learned and predicted simultaneously.

The example of the Multi-task Learning model is illustrated in figure 5. For the aspect detection task, the output layer is a dense layer with 8 nodes, corresponding to 8 aspects. The sigmoid activation is used because this task is multi-label classification and a review can have more than one aspect. For the sentiment classification tasks, there are 7 tasks corresponding to 7 aspects that are needed to classify the sentiment (except the aspect others). In each task, the output has 4 nodes indicating positive, neutral, negative, or nan (the aspect is absent in this review so there is no sentiment for this aspect). The softmax

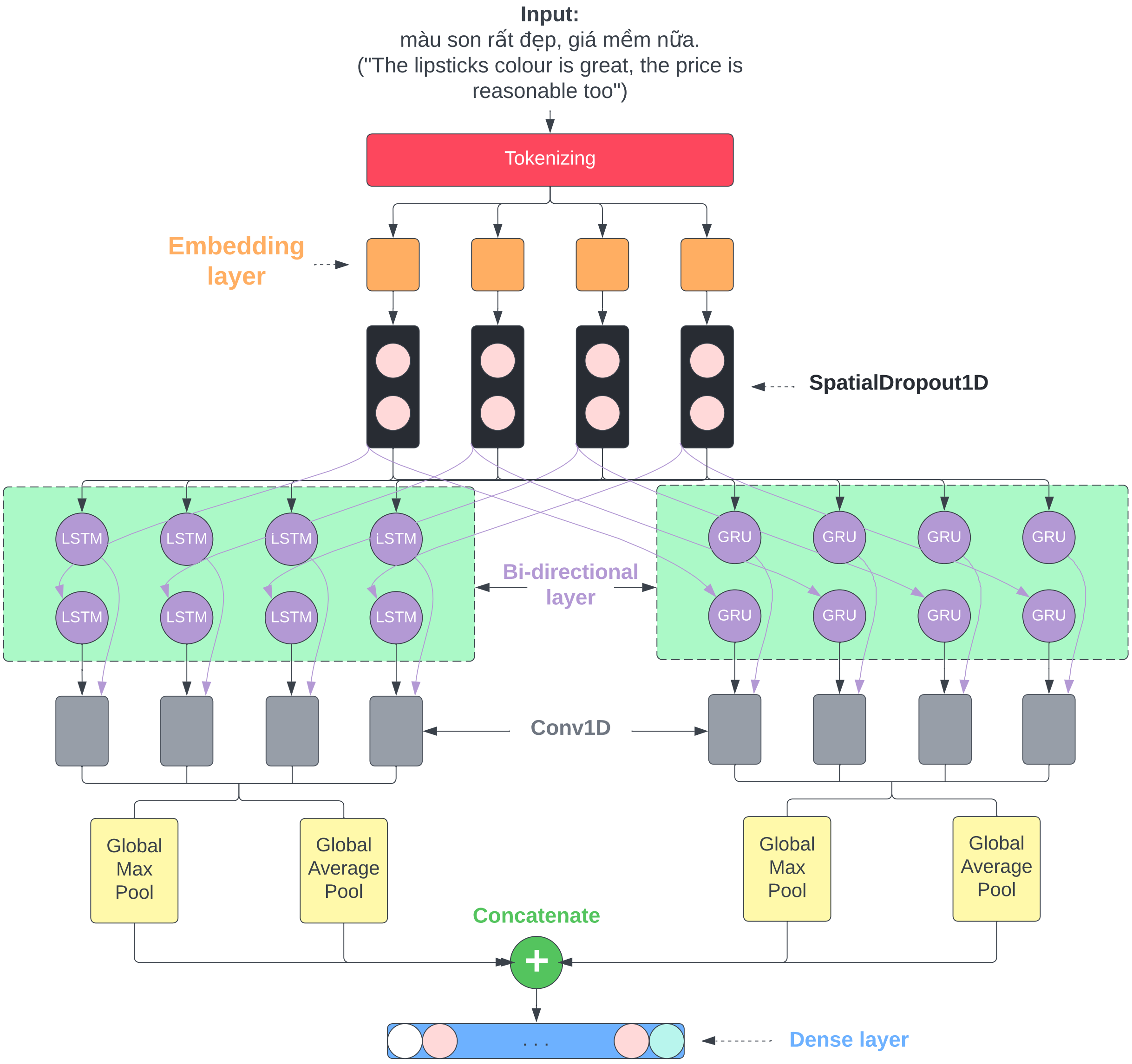


Figure 3: The proposed input and hidden layers

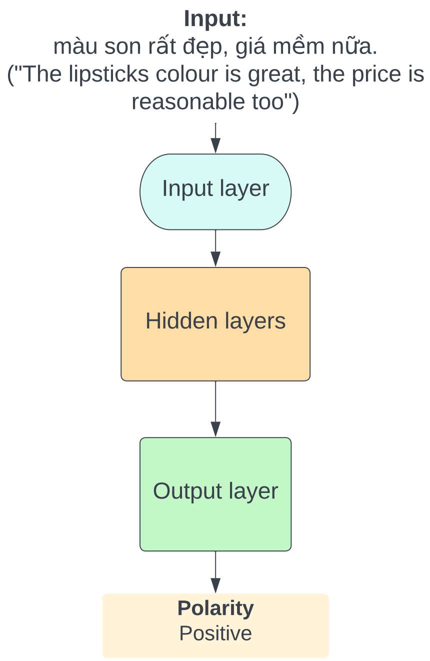

Input:
màu son rất đẹp, giá mềm nữa.
("The lipsticks colour is great, the price is reasonable too")
Input layer
Hidden layers
Output layer
Polarity
Positive

Figure 4: The example STL model for classifying the sentiment of aspect COLOUR

activation function is used for these tasks.

5 Experiment

5.1 Experimental settings

The embedding layer has the dimension of 1024, which is also the dimension of the ELMO pre-trained word embedding. The number of units in BiLSTM or BiGRU is 400 units and the activation function in these layers is tanh function. In the Conv1D layer, the kernel size is 2 and the filter is 128, which means reducing the input dimension from 400 to 128.

We use binary_crossentropy loss function for the aspect detection task and categorical_crossentropy for the sentiment classification tasks. The Adam (Kingma & Ba, 2014) is used as the optimizer with the learning rate at 0.0001. The batch size is 128 and the number of epochs for training is 50 epochs. Early Stopping is used to reduce the overfitting problem. For the implemented code, one can see it at this repository at Github: `https://github.com/linh222/absa_vn_lipsticks_review`.

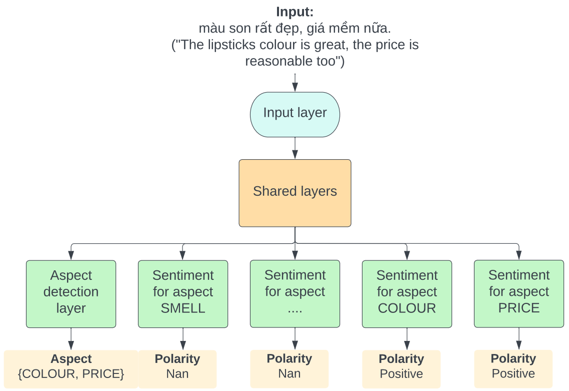

Input:
màu son rất đẹp, giá mềm nữa.
("The lipsticks colour is great, the price is reasonable too")
Input layer
Shared layers
Aspect detection layer
Sentiment for aspect SMELL
Sentiment for aspect ....
Sentiment for aspect COLOUR
Sentiment for aspect PRICE
Aspect
{COLOUR, PRICE}
Polarity
Nan
Polarity
Nan
Polarity
Positive
Polarity
Positive

Figure 5: The example MTL model

5.2 Evaluation metrics

Because of the imbalance of aspect and sentiment in the experimental dataset, F1-score is used to evaluate the performance of models to take into account the imbalance and give a proper view of the effectiveness of the predictive models. The $F1_{ad}$ and $F1_{sc}$ denote for F1-score in aspect detection and sentiment classification tasks. The formula of F1-score (according to (Sokolova & Lapalme, 2009)) is as follow with n is the number of samples and TP, FP, and FN denote for TruePosituve, FalsePositive, and FalseNegative, respectively:

$$Precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FP_i}$$

$$Recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FN_i}$$

$$F1 - score = \frac{2(Precision * Recall)}{(Precision + Recall)}$$

5.3 Results and Discussion

Weighted averages of F1-Score are used to measure the performance of our models. The table 4 illustrates the performance of implemented model architectures on the dataset. Most models achieve a performance higher than 87% for single-task learning and over 80% for multi-task learning. The best model in both approaches is BiGRU + Conv1D with 98.09% F1_ad, 91.01% F1_se for the single-task learning approach, and 97.1% F1_ad, 86.92% F1_se for the multi-task learning approach.

Model	STL		MTL	
Metric	F1_ad	F1_sc	F1_ad	F1_sc
BiLSTM	97.62	87.56	94.12	81.44
BiGRU	97.90	87.22	95.61	80.94
BiLSTM +Conv1D	98.00	89.90	96.89	86.26
BiGRU +Conv1D	**98.09**	**91.01**	**97.51**	**86.92**
BiLSTM +BiGRU +Conv1D	97.92	90.15	96.63	85.65

Table 4: Results of implemented models on the dataset (%)

Aspect	STL	MTL
SMELL	99.23	98.01
COLOUR	97.06	96.60
STAYINGPOWER	98.48	95.67
PRICE	98.33	96.37
SHIPPING	97.69	97.15
PACKING	98.18	95.89
TEXTURE	96.80	94.88
OTHERS	97.48	94.68

Table 5: The F1-score for aspect detection in each aspect(%)

Aspect	Positive	Neutral	Negative
SMELL	94.22	0.00	74.01
COLOUR	95.66	36.90	67.48
STAYING-POWER	84.97	47.63	86.49
PRICE	99.01	60.00	0.00
SHIPPING	93.92	33.23	91.97
PACKING	98.55	0.00	44.90
TEXTURE	92.22	53.87	77.57

Table 6: The F1-score of sentiment in each aspect in Single-Task Learning(%)

Aspect	Positive	Neutral	Negative
SMELL	93.41	0.00	76.32
COLOUR	94.35	28.04	63.95
STAYING-POWER	82.98	27.84	89.73
PRICE	96.32	0.00	0.00
SHIPPING	92.01	33.28	90.42
PACKING	94.28	0.00	17.83
TEXTURE	89.52	48.92	76.90

Table 7: The F1-score of sentiment in each aspect in Multi-Task Learning(%)

Table 5 proves that our models are robust for aspect detection task. The F1-score for the aspect detection task is always higher than 96% for the single-task learning approach and 94% for the multi-task learning approach.

For sentiment classification task, the results on table 6 and 7 show that the performance in classifying the sentiment Positive is better than other sentiments. The model can detect the sentiment of some aspects such as STAYINGPOWER, SHIPPING, TEXTURE very accurately. However, some other aspects are poorly in sentiment classification such as SMELL, PACKING, especially on the sentiment Neutral and Negative. The reason is there are a lot of positive reviews for beauty products on Shopee while very few neutral and negative reviews. The imbalance of the dataset is one of the big challenges which will be addressed in future work.

6 Conclusion and Future Work

This paper deals with the aspect-based sentiment analysis of beauty products reviews on e-commerce websites. In this paper, we presented a new dataset containing 16,277 reviews about lipstick in e-commerce platforms for the task aspect-based sentiment analysis. There are 32,775 pairs of aspect and sentiment in the dataset. For the task of predicting the aspect and sentiment of reviews, we compared single-task learning and multi-task learning and received the result that single-task learning is better than multi-task learning. However, the implementation and complexity of single-task learning are significantly higher than multi-task learning so this is a trade-off between accuracy and complexity. For model architecture, the combination of BiGRU and Conv1D outperformed other model architecture in both single-task learning and multi-task learning. The best F1-score belonged to BiGRU+Conv1D in the single-task learning approach at 98.09% for aspect detection and 91.01% for sentiment classification.

For future work, we are considering building an automatic pipeline to collect reviews in e-commerce platforms, process the reviews, predict the aspect and sentiment of reviews and visualize the results on a dashboard. This pipeline will bring a broader look for sellers to understand their products and for customers to consider before making a purchase. In addition, we will apply some state-of-the-art models such as transformer models to improve the accuracy of sentiment classification.

References

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling.* arXiv. doi: 10.48550/ARXIV.1412.3555

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104

He, R., Lee, W., Ng, H., & Dahlmeier, D. (2019, 01). An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In (p. 504-515). doi: 10.18653/v1/P19-1048

Hochreiter, S., & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.* arXiv. doi: 10.48550/ARXIV.1412.6980

Luo, H., Li, T., Liu, B., & Zhang, J. (2019). *Doer: Dual cross-shared rnn for aspect term-polarity co-extraction.* arXiv. doi: 10.48550/ARXIV .1906.01794

Nguyen, H. T., Nguyen, H. V., Ngo, Q. T., Vu, L. X., Tran, V. M., Ngo, B. X., & Le, C. A. (2018). Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, *34*(4), 295–310.

Nguyen, K. T.-T., Huynh, S. K., Phan, L. L., Pham, P. H., Nguyen, D.-V., & Van Nguyen, K. (2021). *Span detection for aspect-based sentiment analysis in vietnamese.* arXiv. doi: 10.48550/ARXIV.2110.07833

Nguyen, K. V., Nguyen, V. D., Nguyen, P. X. V., Truong, T. T. H., & Nguyen, N. L.-T. (2018). Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th international conference on knowledge and systems engineering (kse)* (p. 19-24). doi: 10.1109/KSE.2018.8573337

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., . . . Eryiğit, G. (2016, January). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *International Workshop on Semantic Evaluation* (p. 19 - 30). San Diego, United States. doi: 10.18653/v1/S16-1002

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 486–495).

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014, August). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 27–35). Dublin, Ireland: Association for Computational Linguistics. doi: 10.3115/v1/S14-2004

Sokolova, M., & Lapalme, G. (2009, 07). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*, 427-437. doi: 10.1016/j.ipm.2009.03.002

Vu, X.-S., Vu, T., Tran, S. N., & Jiang, L. (2019). Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of the international conference recent advances in natural language processing (ranlp).*

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606–615). Austin, Texas: Association for Computational

Linguistics. doi: 10.18653/v1/D16-1058

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). *A survey on aspect-based sentiment analysis: Tasks, methods, and challenges.* arXiv. doi: 10.48550/ARXIV.2203.01054

Zhou, X., Wan, X., & Xiao, J. (2015). Representation learning for aspect category detection in online reviews. In *Proceedings of the twenty-ninth aaai conference on artificial intelligence* (p. 417–423). AAAI Press.

Intent Detection and Slot Filling from Dependency Parsing Perspective: A Case Study in Vietnamese

Phu-Thinh Pham, Duy Vu-Tran, Duc Do, An-Vinh Luong, Dien Dinh
University of Science, Ho Chi Minh city, Vietnam
Vietnam National University, Ho Chi Minh city, Vietnam
{phpthinh18, vtduy18}@apcs.fitus.edu.vn
{dotrananhduc, anvinhluong}@gmail.com
ddien@fit.hcmus.edu.vn

Abstract

Spoken language understanding (SLU) systems using deep learning techniques require effective intent detection and slot filling models. Previous studies in this field taking advantage of sequence to sequence models have achieved good results. However, they focus on the locality of words and thus are sensitive to surrounding terms. In this paper, we introduce a new approach for this problem inspired by the dependency parsing techniques via a biaffine model to give the system a global view of the input. The experiments on PhoATIS dataset for Vietnamese have shown that our joint model for intent detection and slot filling obtains potential results.

1 Introduction

Spoken language understanding (SLU) has been applied to many chatbot applications in recent years. Intent detection and slot filling are two main tasks in this field for building task-oriented dialog systems. The purpose of intent detection task is to classify users' intent and that of slot filling task is to extract semantic constituents from the natural language utterances (Tur and De Mori, 2011). The most common approach for intent detection task is using a classifier based on [CLS] context representation. In parallel, the slot filling task is usually considered as a sequence to sequence problem, with the help of conditional random fields (CRFs) and recurrent neural network (RNN). Normally, these two tasks are considered as two distinct tasks, thus implemented separately, although the slots intuitively depend on the intent (Goo et al., 2018). Hence, some studies have proposed joint models based on the correlation between two tasks, enhancing the performance of each other (Goo et al., 2018; Dao et al., 2021; Wu et al., 2020; Wang et al., 2018; Chen et al., 2019).

Briefly summarized, most of the previous studies take advantage of autoregressive model or sequence to sequence architecture to solve the problems (Wu et al., 2020). For example, conditional random field (CRF) is a common approach for slot filling tasks since it considers the correlations between tags. However, we argue that this approach heavily relies on the locality of words and we need to provide the model with a global view of the input.

In this study, we propose a joint model for intent detection and slot filling tasks inspired by the dependency parsing technique. For slot filling task, we reformulate it as the task of identifying the span of a slot and assigning its category, following the study of Yu et al. (2020). In parallel, we consider the intent detection task as the task of classifying the intent labels of the span from the beginning to the end of an utterance. Furthermore, we incorporate the intent context information with an intent-slot attention layer into slot filling, following Dao et al. (2021). Our system uses two biaffine modules (Dozat and Manning, 2016) for the two tasks to estimate the scores to all spans in an utterance. After that, the logits are decoded to return the final results to satisfy the constraints.

We evaluate our system on the PhoATIS dataset (Dao et al., 2021), the first public dataset for Vietnamese intent detection and slot filling. In spite of being the 17th most spoken language in the world (Eberhard et al., 2019), the research attention in this field for Vietnamese has not gained any consideration until the appearance of PhoATIS. The experiments show that our system achieved competitive results and set a new benchmark for this corpus and this language.

In summary, we: (1) introduce a new approach for intent detection and slot filling system inspired by the graph-based dependency parsing technique; (2) propose a joint model that obtains better performance on the Vietnamese dataset.

2 Related work

The introduction of ATIS dataset (Hemphill et al., 1990) has motivated research studies in natural langugae understanding (NLU) and there are efforts to conquer this field. (Chen et al., 2019) has explored the influence of BERT (Devlin et al., 2018) on SLU systems by proposing a joint intent classification and slot filling model based on BERT. With the help of such powerful architecture, they obtain significant improvement in intent classification accuracy, slot F1 score, and sentence-level semantic frame accuracy.

For Vietnamese, PhoATIS (Dao et al., 2021) has been introduced as the first public intent detection and slot filling dataset, setting a starting point for future Vietnamese SLU research. In addition, they also propose a joint model based on the work of Devlin et al. (2018), extending the model by integrating an intent-context attention layer. It helps the model to recognize slots in an utterance more effectively with intent context information. With this architecture, they achieve potential results on the Vietnamese dataset, significantly outperforming the original work.

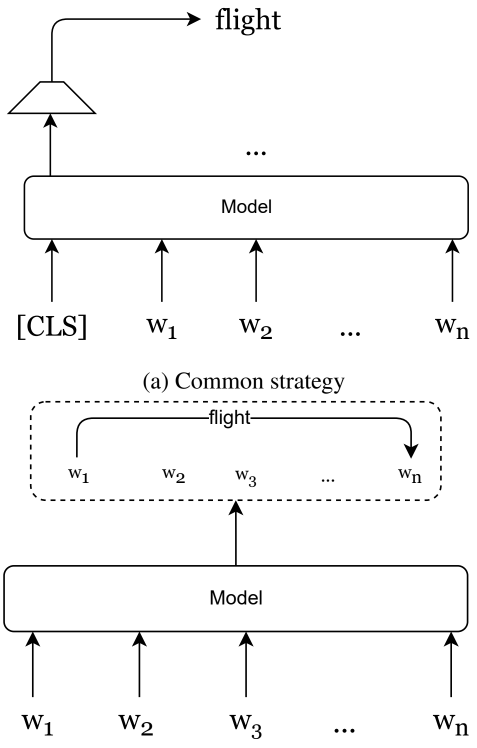
The study of Yu et al. (2020) gives a novel view to named entity recognition (NER) task, as well as sequence labeling problems in general, by applying the ideas from graph-based dependency parsing. It uses a biaffine model (Dozat and Manning, 2016) to score all possible spans in a sentence, enabling the model to predict named entities more accurately. From scores of all pairs of start and end tokens, it ranks the candidate spans based on their scores and selects top-ranked spans satisfying the constraints for flat or nested entities. The experimental results show that the model can handle nested entities well and gain competitive performance on both nested and flat NER.

3 Method

In this section, we first briefly introduce our novel approach for both intent detection and slot filling tasks. Thus, we describe the proposed joint model based on the dependency parsing technique.

3.1 Intent detection

In general, the common strategy for the intent detection task is to predict the intent based on the hidden state of the first special token ([CLS]). In this paper, we reformulate it as the task of classifying the whole sentence, represented by a span from the beginning to the end of the sentence. Figure 1 compares our proposed approach and the common approach.



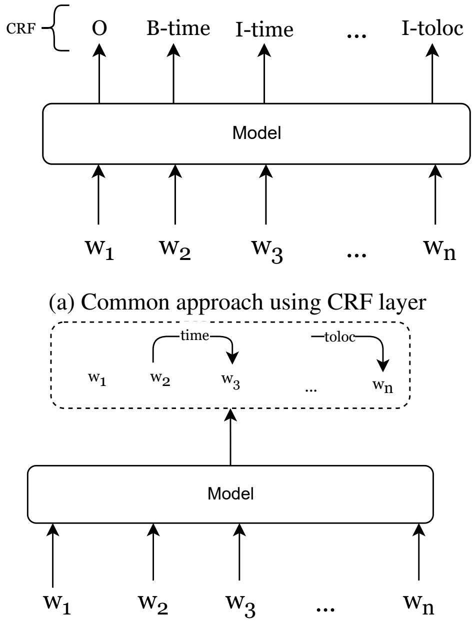
(a) Common strategy

(b) Our new approach

Figure 1: Comparison between the common method and our proposed approach

3.2 Slot filling

Previous studies consider slot filling task as a sequence labeling problem, with a CRF layer for prediction. In our approach, to provide a more general view, we adopt ideas from the graph-based dependency parsing model inspired by the study of Yu et al. (2020). In detail, we reformulate slot filling as the task of identifying the start and end indices of a slot, as well as classifying its category. Table 2 illustrates the difference between our approach and the previous approach. By using the biaffine model of Dozat and Manning (2016), our model scores all possible spans that could form a slot in an utterance. Thus, our system ranks these spans based on the logits predicted by the biaffine model and accordingly selects top-ranked spans complying with constraints that no two slots are overlapped. Formally, given an $n \times n \times c$ tensor T from our model, where n is the length of the utterance and c is the number of slot types $+1$ (for non-slot), each span i with the start and end indices $s_i \leq e_i$ is assigned the category c with the highest



(a) Common approach using CRF layer

(b) Our approach using Biaffine layer

Figure 2: Comparison between previous and our approaches

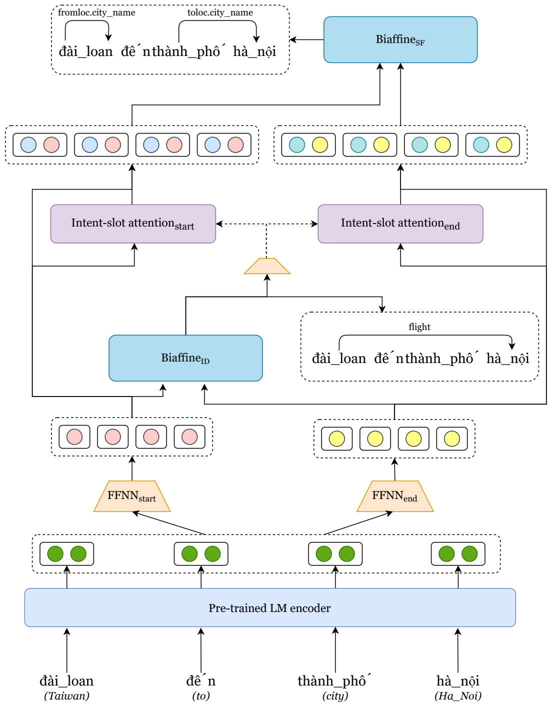


Figure 3: Illustration of our proposed model

score:

$$y'(s_i, e_i) = \arg\max_c T(s_i, e_i, c) \tag{1}$$

Finally, all spans whose category is different from non-slot are ranked based on their scores in descending order. A slot i will be selected if there is no higher-ranked slot j such that $s_i \leq s_j \leq e_i$ or $s_j \leq s_i \leq e_j$.

3.3 Model architecture

The architecture of our joint model is illustrated in Figure 3, consisting of 7 layers: an encoding layer, two feed-forward neural network (FFNN) layers, two intent-slot attention layers, and two biaffine layers.

Encoding layer

In the encoding layer, we employ a pre-trained Transformer-based language model (LM) to generate context-dependent sentence representations of an utterance. Here, we utilize XLM-R (Conneau et al., 2019) as the encoder for the syllable-level dataset and PhoBERT (Nguyen and Nguyen, 2020) for its automatically word-segmented variant. Given an n-length input token sequence $\boldsymbol{w} = (w_1, w_2, ..., w_n)$, the output produced by the encoding layer is feature embeddings $\mathbf{c}_i$ representing the i^{th} token.

FFNN layers

Following the encoding layer, two separate FFNNs are used to extract different representations for the start and end of the spans. This allows the model to distinguish different contexts of start and end of spans and reduce the dimensions of the encoder's output. In particular, each layer feeds $\mathbf{c}_i$ into a single-layer feed-forward network to obtain the start or end representation of token i:

$$\mathbf{v}_i^{start} = \text{FFNN}_{start}(\mathbf{c}_i) \tag{2}$$

$$\mathbf{v}_i^{end} = \text{FFNN}_{end}(\mathbf{c}_i) \tag{3}$$

Intent-slot attention layers

Following the architecture presented in (Dao et al., 2021), we use an attention mechanism to take advantage of intent information for better slot filling performance. In particular, each intent-slot attention layer takes the input from the start/end (s/e) FFNN layer and the probability vector $\mathbf{p}$ from the output of the intent detection module to produce s/e intent-specific vectors. In details, the layer creates an intent label embedding $\mathbf{r}$ via a label weight matrix $\mathbf{R}$ and uses it to give the intent-specific vector $\mathbf{h}_i$:

$$\mathbf{r}^{s/e} = \mathbf{R}^{s/e}\mathbf{p} \tag{4}$$

$$\alpha_i^{s/e} = \frac{\exp((\mathbf{r}^{s/e})^T \mathbf{v}_i^{s/e})}{\sum_{j=1}^{n} \exp((\mathbf{r}^{s/e})^T \mathbf{v}_j^{s/e})} \tag{5}$$

$$\mathbf{h}_i^{s/e} = \alpha_i^{s/e}\mathbf{r}^{s/e} \tag{6}$$

After that, a sequence of vectors $\mathbf{s}_{1:n}$ is created, where s_i is the concatenation of intent-specific vector and the corresponding start/end representation from the FFNN layer:

$$\mathbf{s}_i^{s/e} = \mathbf{h}_i^{s/e} \circ \mathbf{v}_i^{s/e} \tag{7}$$

Biaffine layers

Our model consists of two biaffine layers of Dozat and Manning (2016), one for intent detection and the other for slot filling task. To be more specific, the layer for intent detection takes two sequences of vectors $\mathbf{v}_{1:n}^{start}$ and $\mathbf{v}_{1:n}^{end}$ as the inputs while the other one feeds $\mathbf{s}_{1:n}^{start}$ and $\mathbf{s}_{1:n}^{end}$. Each layer returns an $n \times n \times c$ tensor, where c is the number of intent labels for intent detection task and the number of slot types $+1$ for slot filling task as explained in 3.2.

3.4 Joint training

The learning objective of our model is to classify the correct intent and correct slot type for each valid span. Therefore, we consider them as two multi-class classification problems and optimize our models for both tasks with softmax cross-entropy. Given the tensor T_{ID} produced by the biaffine layer for intent detection, the probability vector $\mathbf{p}$ is calculated via a softmax function:

$$p_i = \frac{\exp(T_{ID}(1, n, i))}{\sum_{j=1}^{k} \exp(T_{ID}(1, n, j))} \tag{8}$$

where k is the number of intent classes. Based on the vector $\mathbf{p}$, a loss $\mathcal{L}_{ID}$ for intent classification is then computed:

$$\mathcal{L}_{ID} = -\sum_{i=1}^{k} y_i \log(p_i) \tag{9}$$

For slot filling, a cross-entropy objective loss $\mathcal{L}_{SF}$ is calculated from the output T_{SF} of biaffine layer:

$$p'(s, e, i) = \frac{\exp(T_{SF}(s, e, i))}{\sum_{j=1}^{c} \exp(T_{SF}(s, e, j))} \tag{10}$$

$$\mathcal{L}_{SF} = -\sum_{s=1}^{n}\sum_{e=s}^{n}\sum_{i=1}^{c} y(s, e, i) \log(p'(s, e, i)) \tag{11}$$

The final loss $\mathcal{L}$ is the weighted sum of the intent detection loss $\mathcal{L}_{ID}$ and slot filling loss $\mathcal{L}_{SF}$.

$$\mathcal{L} = \delta\mathcal{L}_{ID} + (1 - \delta)\mathcal{L}_{SF} \tag{12}$$

where $0 < \delta < 1$ is the mixture weight.

Model	**Intent**	**Slot**	**Sent.**
Syllable-level			
JointBERT+CRF	97.42	94.62	85.39
JointIDSF	97.56	94.95	86.17
Our model	97.61	**95.05**	85.89
Word-level			
JointBERT+CRF	97.40	94.75	85.55
JointIDSF	97.62	94.98	86.25
Our model	97.80	**95.43**	**87.05**

Table 1: Results on the test set. Numbers written in bold indicate that the improvement of our model is statistically significant with p-$value$ < 0.05 under t-test.

4 Experiments and Results

4.1 Experimental setup

We evaluate our models on the PhoATIS dataset (Dao et al., 2021) and conduct the experiments on both word and syllable levels. The dataset consists of 4478, 500 and 893 utterances for train, validation and test set, respectively with 28 intent labels and 82 slot types. For hyper-parameters, we follow the same configuration in the original work of Dao et al. (2021). To optimize the model, we use AdamW optimizer (Loshchilov and Hutter, 2017) and test on different δ in $\{0.05, 0.1, 0.15, ..., 0.95\}$ to select the optimal value. The batch size is set to 32 and the number of Transformer layers, attention heads and hidden sizes are 12, 12 and 768 respectively.

The metrics used for evaluation are the intent accuracy for intent detection, the F_1-score for slot filling and the overall sentence accuracy (Louvan and Magnini, 2020; Weld et al., 2021). During training, we compute the average score of intent accuracy and F_1 score at each epoch to select the checkpoint achieving the best performance on the validation set. We train the model for 100 epochs with the early stopping strategy. All results are reported on average over 3 runs with 3 different random seeds.

4.2 Results

Table 1 gives information about the results on the test set of our models, in comparison to the baseline JointBERT+CRF and JointIDSF reported in (Dao et al., 2021). Since we evaluate our models using the syllable-level dataset and its word-segmented variant, the results are presented in two comparable settings.

In syllable level, our model achieves 97.61%, 95.05% and 85.89% for intent accuracy, slot F_1

score and sentence accuracy, respectively. Especially, the slot F_1 score improvement over JointIDSF is statistically significant with $p\text{-}value < 0.05$. On the other hand, our model obtains better results in word level, with 97.80%, 95.43% and 87.05% for intent accuracy, slot F_1 and sentence accuracy respectively. In comparison to JointIDSF baseline, the slot F_1 score and sentence accuracy are statistically significant with $p\text{-}value < 0.05$.

From the results, we find that our models achieve better performance, except for the sentence accuracy on the syllable-level dataset. This can be explained by the fact that representing Vietnamese tokens at the syllable level cannot capture the whole meaning compared to word-segmented tokens. Thus, employing such information-lost token representations to compute the scores for all possible spans has a negative impact on model performances, leading to low sentence accuracy although the intent accuracy and slot F_1 are higher than the JointIDSF baseline. The significant difference in the sentence accuracy between the syllable-level dataset and its automatically word-segmented variant, 85.89% and 87.05% respectively, is the strong evidence for our explanation.

4.3 Ablation study

To evaluate the effectiveness of individual components in our system, we do an ablation study using the word-level setup because of its better performance. In particular, we remove selected components in our model and train them for evaluation.

To verify the contribution of two intent-slot attention layers in our proposed architecture, we sequentially remove one and then both of them. With only one attention layer, our system creates the s/e intent-specific vectors by sharing common parameters (using $\mathbf{r} = \mathbf{Rp}$ in equation 4 and replacing $\mathbf{r}^{s/e}$ in equations 5 and 6 by $\mathbf{r}$). Meanwhile, when two attention layers are removed, our model becomes a joint model consisting of two biaffine modules with the same s/e representations (using $\mathbf{s}_i^{s/e} = \mathbf{v}_i^{s/e}$ in equation 7). Finally, to confirm the influence of our new approach for intent detection task, we replace the biaffine layer responsible for classifying intents by a linear prediction layer using the [CLS] token.

Table 2 clearly shows that removing any components from our full model has a negative impact on its performance in all three metrics. When we completely remove the intent-slot attention layers, the performance witnesses a significant drop by 2.36% in sentence accuracy (from 87.05% to 84.69%). Adding an attention layer helps our model improve 1.8% score from 84.69% to 86.49%, 0.56% lower than the full model, clearly proving the contribution of this component in our architecture. Besides, when we replace the biaffine layer for intent detection with a single-layer feed-forward network based on the contextualized embedding of the classification token [CLS], the performance of our full model is reduced by 1.05% to 86.00%.

	Intent	**Slot**	**Sent.**
Our model	97.80	95.43	87.05
- One attention	97.76	95.23	86.49
- No attention	97.46	94.84	84.69
- [CLS] token	97.65	95.10	86.00

Table 2: Ablation study results on the test set.

5 Conclusion

In this paper, we have presented our work for Vietnamese intent detection and slot filling tasks. By proposing an effective architecture for jointly training intent detection and slot filling, we achieve better performance than the previous work JointIDSF. In particular, we adopt the ideas and techniques from dependency parsing to apply to our models, along with taking advantage of the intent-slot attention layer to integrate intent context information for better slot filling. In addition, we find that our proposed architecture works better at the word level compared to the syllable level. Furthermore, we empirically conduct experiments on the dataset to verify the contribution of each component in the architecture.

Acknowledgments

This research is supported by research funding from Advanced Program in Computer Science, University of Science, Vietnam National University - Ho Chi Minh City.

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent detection and slot filling for vietnamese. *arXiv preprint arXiv:2104.02021*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

David Eberhard, Gary Simons, and Chuck Fennig. 2019. *Ethnologue: Languages of the World, 22nd Edition*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *arXiv preprint arXiv:2011.00564*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235*.

HENRY Weld, Xiaoqi Huang, SIQU Long, Josiah Poon, and SOYEON CAREN Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding. *arXiv preprint arXiv:2101.08091*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv preprint arXiv:2010.02693*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.

Annotating Entity and Causal Relationships on Japanese Vehicle Recall Information

Hsuan-Yu Kuo, Youmi Ma, and Naoaki Okazaki

Tokyo Institute of Technology,
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550 Japan
{kuo.hsuanyu,youmi.ma}@nlp.c.titech.ac.jp
okazaki@c.titech.ac.jp

Abstract

A vehicle recall system is a process of recalling and repairing vehicles with defective designs or potential for accidents and failures. The recall document concisely explains the circumstances and causes of product defects. This paper presents two types of annotations on public vehicle recall reports, part entities and their relations, and causality. We annotated 6,394 car-recall text documents. Named entity and relation annotation suggests a relationship between the elements of an automobile, and causality annotation indicates the cause of a malfunction. The entity and relation annotation and causality annotation allow the system to automatically extract knowledge in the automotive design domain. Subsequently, we present the experimental results for named entity recognition and relation extraction and causality extraction of our annotated corpus to verify the feasibility of building a system for extracting part information and causality. Finally, the experimental results show that employing named entity and relation information as the external knowledge improves causality extraction.

1 Introduction

A defect/bug in a product causes significant losses and damages to both users and manufacturers. Although manufacturers conduct design/code reviews to ensure the quality of a product, manual reviews involve various challenges, such as correctness, comprehensiveness, cost, and development of human experts. Therefore, we expect computers to automate or assist in the review process.

Information extraction from unstructured text is a straightforward approach for computers to learn expert knowledge, and researchers have applied information extraction in various fields such as news (Chinchor, 1998), biomedical (, 2002), clinical (Demner-Fushman et al., 2009; Rumshisky et al., 2016), and business (Bahja, 2020). However, no previous work has explored its applications in the manufacturing industry.

In this study, we explore scenarios for information extraction in the car industry. As the design and review records of each company in the car industry are kept strictly confidential, we cannot share a corpus and dataset created for this domain. Instead, we focus on vehicle recall reports published by the Japanese government.

A recall system is a system in which an automobile manufacturer, at its discretion, notifies the Minister of Land, Infrastructure, Transport and Tourism (MLIT)[1] in advance of a recall or repair of a product due to a problem in the design or manufacturing process to prevent further accidents and problems. The text describing the situation of each recall is available on the MLIT website[2]. A recall text briefly describes the circumstances and causes of product defects, possibly useful for extracting information from design reviews in the manufacturing process.

Useful information in a car recall text includes entity mentions, entity relations, and causal relationships. For example, consider the following sentence: “Due to inappropriate electrical circuitry in the aux-

[1] https://www.mlit.go.jp/en/

[2] https://www.mlit.go.jp/jidosha/recall.html

iliary braking device (electromagnetic retarder), the braking light does not turn on when the electromagnetic retarder is activated." There are entity mentions (e.g., "auxiliary braking device," "electromagnetic retarder," and "braking light") and entity relations (e.g., "electromagnetic retarder" *is an* "auxiliary braking device"; "braking light" *is connected with* "electromagnetic retarder"). The text also contains a causal relationship between "inappropriate electrical circuitry in the auxiliary braking device" and "the braking light does not come on when the electromagnetic retarder is activated." These causal relationships are extremely useful for specifying the reason for a malfunction, thus helping to avoid related problems in product design or reviews.

Recognizing causality requires knowledge of individual components and their relations in vehicles. In addition, relation instances extracted from recall information can be used to build a knowledge base (KB) for manufacturing cars.

In this study, we build a corpus of Japanese vehicle recall information, where 6,394 documents are annotated with named entities (NEs), their relations, and causal relations to build a system that assists the review process when designing and manufacturing vehicles. To verify the feasibility of building a system for extracting part information and causality, we employ a joint NER/RE model to extract information from the annotated corpus. The main contributions of this paper can be summarized as follows:

- According to our research, this is the first work annotating NEs, relations, and causalities on vehicle recall information[3].
- We report the experimental results on named entity recognition, relation extraction, and causality extraction. We also show that incorporating knowledge about entities and their relations improves the performance of causality extraction.
- We summarize issues in building the corpus, hoping that these findings will be useful for building corpora in other manufacturing fields.

2 Related work

Considering that the ultimate goal of this study was to assist the reviewing process of product design, our goal was to build a model and KB to infer possible defects in a given design. Therefore, research on causality extraction was the most relevant to our study. A common approach for causality extraction is to build an annotated corpus and train a model on the corpus. In this section, we describe existing corpora for causality extraction in general and specific domains.

SemEval 2007 Task 4 (Girju et al. , 2007) considered the task of recognizing a semantic relation (including a cause-effect relation) between simple nominals as a binary classification problem. SemEval 2010 Task 8 (Hendrickx et al., 2010), a direct successor of SemEval 2007 Task 4, also addressed the same task but formalized the task as a multiclass classification problem. The datasets of these two tasks use Wikipedia as the source documents.

BECauSE 1.0 (Dunietz et al., 2015) annotated causality instances in the New York Times (NYT) corpus (Sandhaus, 2008). BECauSE 2.0 (Dunietz et al., 2007), a successor of BECauSE 1.0, includes relations overlapping with causality. In addition, CaTeRS (Mostafazadeh et al., 2016) is an annotation scheme that captures a set of temporal and causal relations between events, and the authors annotated a total of 1600 sentences sampled from ROCStories (Mostafazadeh et al., 2016). Inspired by TimeML (Pustejovsky et al., 2003), Mirza et al. (Mirza et al., 2014) proposed guidelines for annotating the causality relation in the TempEval-3 corpus and a rule-based algorithm for automatic annotation.

Biomedical literature is the most explored domain for causality extraction. BioInfer (Pyysalo et al., 2007) presents an annotation scheme and corpus capturing NEs and their relationships, along with a dependency analysis of a sentence. BioCause (Mihuailua et al., 2013) is an annotated corpus with open-access full-text biomedical journal articles belonging to the subdomain of infectious diseases. The corpus annotates linguistic causality instances consisting of a causal trigger (usually a connective), cause, and effect. Using the defined scheme, the researchers added 851 casual relations annotations

[3] We will release the corpus to the public after this paper is accepted.

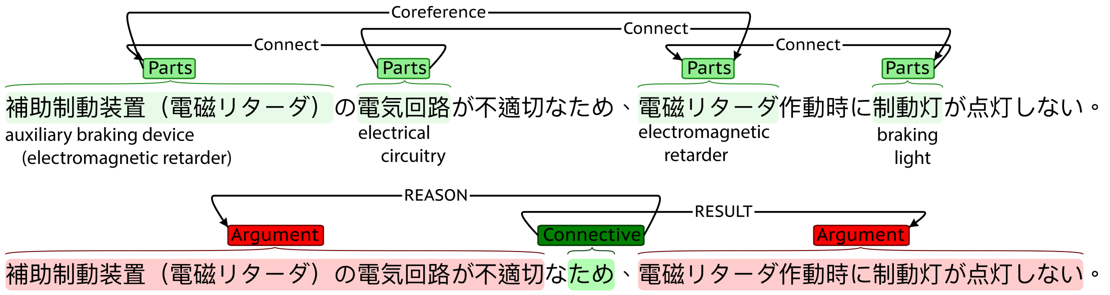

Coreference
Connect
Connect
Connect
Parts
Parts
Parts
Parts
補助制動装置（電磁リターダ）の電気回路が不適切なため、電磁リターダ作動時に制動灯が点灯しない。
auxiliary braking device
(electromagnetic retarder)
electrical
circuitry
electromagnetic
retarder
braking
light
REASON
RESULT
Argument
Connective
Argument
補助制動装置（電磁リターダ）の電気回路が不適切なため、電磁リターダ作動時に制動灯が点灯しない。

Figure 1: Upper figure shows an example document annotated with NEs and relations. Lower figure shows the same document annotated with causality. BRAT (Stenetorp et al., 2012) is used for annotation and visualization. Translation: Due to inappropriate electrical circuitry in the auxiliary braking device (electromagnetic retarder), the braking light does not come on when the electromagnetic retarder is activated.

to the collection of articles. The corpus is pre-annotated with NEs and events of genes and their interactions (e.g., positive and negative regulations).

Inspired by efforts in the biomedical domain, this study assumes that interactions between car components refer to the causality chain of a malfunction; therefore, we annotate causal relations on top of NEs and their relationships in the car recall text.

3 Corpus

We explain the target text in Section 3.1, followed by the annotations of NEs and their relations in Section 3.2 and causal annotations in Section 3.3, respectively.

3.1 Data

We crawled 6,394 Japanese documents reporting car recall information from the MLIT website[4]. The average length of documents is approximately 135 letters, 1.7 sentences, and 74 tokens after the tokenization using the Japanese tokenizer MeCab (Kudo, 2006).

3.2 NEs and relations

We define a single entity type PART and five relation types, CONTACT, CONNECT, PART-WHOLE, COREFERENCE, ONEWAY-COREFERENCE, between the two parts. The upper figure of Figure 1 illustrates an example of a document annotated with entities and relations.

3.2.1 Entity type

This study uses a single entity type PART to annotate car parts (components). We do not distinguish between the granularity of car parts (e.g., “cylinder head” and “engine”) and semantic differences of mentions (e.g., “oil filler” as a car component or as a location in a car). We also include the names of car models and other necessary components as PART, but exclude the following text spans:

- A part that cannot be interpreted as a component but only stands for a specific location, for example, “joint section”.
- Design of structures and methods, for example, “water immersion prevention structure” and “four-wheel-drive”.
- Air, for example, “put *air* into a tire”. Similarly, we exclude “electricity” from the annotations.
- A modifying clause of a part entity. For example, we only annotate “program” as PART entity in “program to calculate the amount of particulate matter deposition.”

3.2.2 Relation types

We define five relation types that frequently appear in recall texts and commit to causal relations.

- CONTACT: Part$_1$ is located next or attached to Part$_2$. This relation is useful for causality extraction because it expresses direct contact between the two parts. Figure 2 illustrates an example of the CONTACT relation: “the steering

[4] https://www.mlit.go.jp/jidosha/recall.html

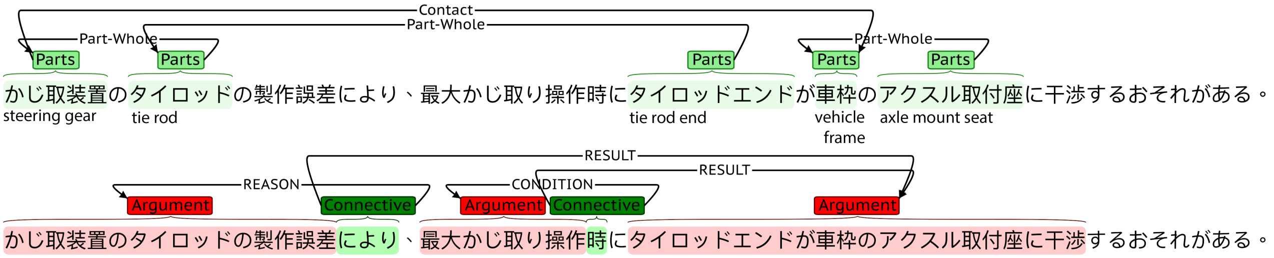

Contact
Part-Whole
Part-Whole
Part-Whole
Parts
Parts
Parts
Parts
Parts
かじ取装置のタイロッドの製作誤差により、最大かじ取り操作時にタイロッドエンドが車枠のアクスル取付座に干渉するおそれがある。
steering gear
tie rod
tie rod end
vehicle frame
axle mount seat
RESULT
RESULT
REASON
CONDITION
Argument
Connective
Argument
Connective
Argument
かじ取装置のタイロッドの製作誤差により、最大かじ取り操作時にタイロッドエンドが車枠のアクスル取付座に干渉するおそれがある。

Figure 2: Translation: Due to a manufacturing error in the tie rod of the steering gear, the tie rod end may interfere with the axle mount seat of the vehicle frame during maximum steering.

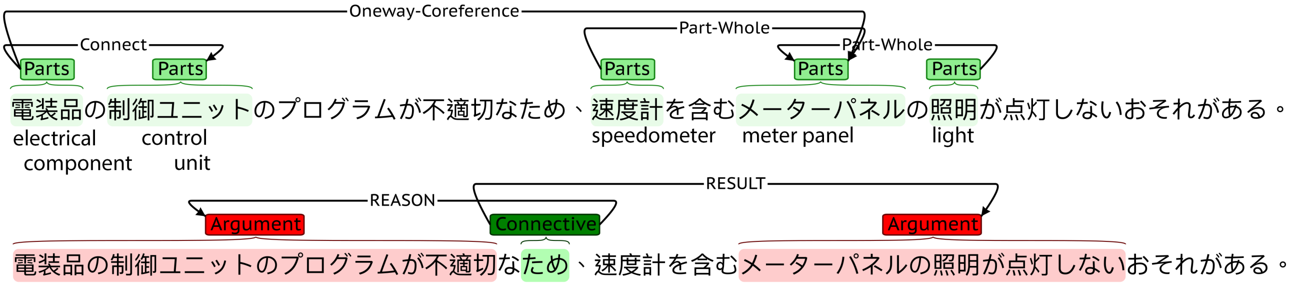

Oneway-Coreference
Part-Whole
Connect
Part-Whole
Parts
Parts
Parts
Parts
Parts
電装品の制御ユニットのプログラムが不適切なため、速度計を含むメーターパネルの照明が点灯しないおそれがある。
electrical component
control unit
speedometer
meter panel
light
RESULT
REASON
Argument
Connective
Argument
電装品の制御ユニットのプログラムが不適切なため、速度計を含むメーターパネルの照明が点灯しないおそれがある。

Figure 3: Example for showing Oneway-Coreference and Part-Whole relations are related to causality.
Translation: Due to an inappropriate program in the control unit of the electrical components, the lights on the meter panel, including the speedometer, may not turn on.

gear" and "the vehicle frame" are in contact with each other. In this example, "a manufacturing error in the tie rod of the steering gear" leads to "the tie rod end may interfere with the axle mount seat of the vehicle frame," indicating that the defect of "the steering gear" can influence "the vehicle frame." In addition, we also annotate *implicit* contact relations between two parts; in other words, contact relations that are not explicitly stated in the text but can be inferred by the context or external knowledge. Consider the example in Figure 2. The CONTACT relation between "the steering gear" and "the vehicle frame" is not explicitly mentioned in the text. If a model for causality extraction is aware of this contact relation, the model can extract causality even without a connective of a causal relation.

- CONNECT: Something connects Part$_1$ and Part$_2$ (electrically, by transfer of matter, or some other forms of transmissions), for example, an "accelerator pedal" is connected to an "engine." Connect relations are also useful clues for recognizing causality relations, expressing an association between the two parts. Consider an example, "the battery is connected with the headlights." If the battery charge is insufficient, we can infer that the headlights are not working. Similar to contact relations, we also annotate the *implicit* connect relations between two parts; for instance, in Figure 1, "the electrical circuity" and "braking light" are connected, but the relation is not explicitly described in the text.

- PART-WHOLE: We consider the following relationships as a part-whole relation:

  1. Part$_1$ is composed by Part$_2$.
  2. Part$_2$ is a part of Part$_1$.
  3. Part$_2$ is spatially contained in Part$_1$.

  The direction of PART-WHOLE is from Part$_2$ to Part$_1$. Figure 2 shows that there is a PART-WHOLE relation from "the tie rod end" to "the tie rod," and the text explains that a manufacturing error in the tie rod of the steering gear leads to the tie rod end possibly interfering with

Entities	Parts	43,158
Relations	Coreference	14,116
	Oneway-Coreference	337
	Part-Whole	13,561
	Contact	8,176
	Connect	4,510
	Total	40,700

Table 1: Statistics of entity and relation annotation.

the axle mount seat of the vehicle frame. In this example, a manufacturing error in the tie rod causes a problem in the tie rod end, showing that PART-WHOLE relations are related to causality.

- COREFERENCE: Part_1 and Part_2 refer to the same *part* entity. This relation is also important for recognizing causal relations where the same entity appears in multiple states and events. The example in Figure 1 illustrates how the COREFERENCE relation can be useful for causality extraction. “The auxiliary braking device (electromagnetic)” and “electromagnetic” refer to the same vehicle *part*, confirming a COREFERENCE relation between them. These two mentions are in clauses indicating the reason and result of a malfunction of the same entity. In the recall report data, clauses containing mentions that refer to the same entity may have causality relations.

- ONEWAY-COREFERENCE: Part_2 refers to Part_1, but Part_1 does not necessarily refer to Part_2 (Part_1 is the subset of Part_2). The direction of ONEWAY-COREFERENCE is from Part_2 to Part_1. In Figure 3, the entity “electrical component” can refer to the “meter panel,” but not vice versa. Similar to the COREFERENCE relation in Figure 1, these two annotated entities are in the clauses explaining the reason and the corresponding result. The reason for extracting ONEWAY-COREFERENCE relations is much the same as that for extracting COREFERENCE.

Table 1 summarizes the statistics of entity and relation annotations for car parts. In total, we annotated 43,158 part entities and 40,700 relations.

Entities	Argument	42,312
	Connective	34,369
	Total	76,681
Relations	REASON	30,098
	RESULT	35,957
	CONDITION	4,510
	Total	70,565

Table 2: Statistics of causality annotation.

3.3 Causality annotation

Following the PDTB 3.0 annotation manual (Webber et al., 2019), we annotated the causality relationships between arguments. We define two argument types: ARGUMENT and CONNECTIVE, and three relation types: REASON, RESULT, and CONDITION. We annotated a causality relation as a combination of REASON and RESULT relations:

$$\begin{aligned} \text{ARGUMENT}_1 &\xleftarrow{\text{REASON}} \text{CONNECTIVE} \\ \text{CONNECTIVE} &\xrightarrow{\text{RESULT}} \text{ARGUMENT}_2 \end{aligned} \tag{1}$$

ARGUMENT_1 presents a cause (reason) of the causal relation, and ARGUMENT_2 presents its effect (result). They are connected to each other by CONNECTIVErelation. Figure 1 shows a real example of a causal relation annotated in a recall document. Here, ARGUMENT_1 is 「補助制動装置（電磁リターダ）の電気回路が不適切な」 “the electrical circuit of the auxiliary braking device (electromagnetic retarder) is inappropriate”; ARGUMENT_2 is 「電磁リターダ作動時に制動灯が点灯しない。」 “braking light does not turn on when the electromagnetic retarder is activated”; CONNECTIVE is 「ため」 “due to” . When a causal relation holds for a specific condition, we also annotate the condition relation:

$$\text{ARGUMENT}_3 \xleftarrow{\text{CONDITION}} \text{CONNECTIVE}$$

Considering the text in Figure 2 as an example, the condition argument is “maximum steering.”

Table 2 summarizes the statistics of causality annotations in the corpus regarding the number of arguments and causal relations.

3.4 Issues during the annotation work

This section describes several issues and ambiguous cases during the annotation work.

3.4.1 No causal connective

In English, a causal relationship is usually expressed with a connective. In contrast, a connective is often dropped in Japanese by simply placing two predicates in the same sentence. In this study, a causal relation is composed of two ARGUMENTS and a CONNECTIVE . A common problem with the above rule is the annotation of connectives. For instance, 「ため」 “due to” or 「場合」“in the case of”, we can simply annotate the connective. However, for examples like 「エンジンが焼き付き走行不能になる」, there is no connective. We annotate the last character of the predicate indicating the reason, 「（焼き付）き」, as a connective to resolve the problem.

3.4.2 Handling of liquids in inter-component relationships

In many cases, it was difficult to determine whether the relation was PART-WHOLE or CONTACT for substances such as fuel and engine oil that go through multiple vehicle parts. In this study, we designed predefined rules to define these relations. Specifically, for the relation between liquid or gas and the tank where it is stored, we annotated the relation as PART-WHOLE. For the relation between the pathways (e.g., fuel pump) and parts where it is used, we annotated the relation as CONTACT .

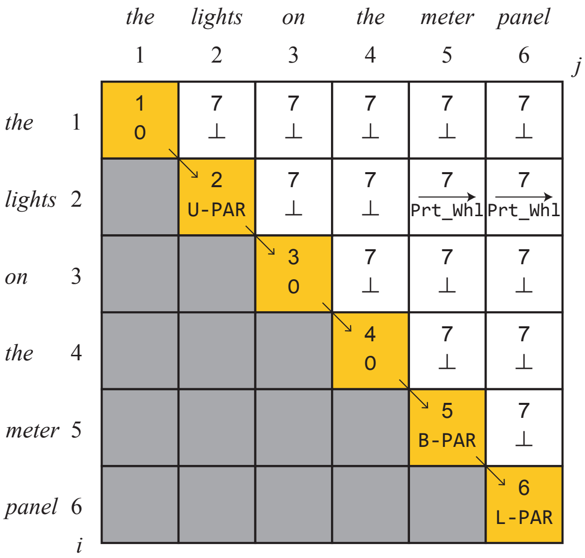
3.4.3 Annotation for materials

In the recall report texts, mentions of material components such as “zinc” sometimes appeared. In this study, we ignore them during entity annotations; however, it may be necessary to annotate material components because they are often involved in the cause of recalls.

4 Experiments

As stated before, corpora are collected to build an automatic information extraction system and KB to promote better and safer vehicle design. In this section, we present experiments conducted to evaluate the extent to which our corpus can help build systems that extract NEs, relations, and causalities. The goals of the experiments are summarized as follows:

1. Evaluate the extent of accuracy of an information extraction model to automatically extract parts information from the text.
2. Evaluate model accuracy on causality extraction.
3. Check if part information helps causality relation extraction.



	the 1	*lights* 2	*on* 3	*the* 4	*meter* 5	*panel* 6
the 1	1 0	7 ⊥	7 ⊥	7 ⊥	7 ⊥	7 ⊥
lights 2		2 U-PAR	7 ⊥	7 ⊥	7 $\overrightarrow{\text{Prt\_Whl}}$	7 $\overrightarrow{\text{Prt\_Whl}}$
on 3			3 0	7 ⊥	7 ⊥	7 ⊥
the 4				4 0	7 ⊥	7 ⊥
meter 5					5 B-PAR	7 ⊥
panel 6						6 L-PAR

Figure 4: Overview of the table filling strategy.

We categorize the causality extraction task into named entity recognition (NER) and relation extraction (RE) tasks. Thus, we use the NER & RE models for part and causal extraction tasks.

4.1 Methodology

To extract NEs and relations jointly, we apply TablERT (Ma et al., 2022), a joint NER/RE model that achieves state-of-the-art performance on the CoNLL04 (Roth & Yih, 2004) and ACE05 English datasets. The model follows the table-filling framework proposed by Miwa & Sasaki (2014) to cast the join extraction of entities and relations as a table-filling problem. As shown in Figure 4, a table is used to represent the label space of both entities and relations. Diagonal cells are entity labels using BILOU notation (Ratinov & Roth, 2009); for example, the entity label for “lights” is *U-Part* meaning “lights” is a unit length PART entity. Upper triangular cells are relation labels, for example, there are PART-WHOLE relations pointing from “lights” to "meter" and "panel". The model fills the table according to the order indicated by the number in each cell. The model first predicts the entity labels (the diagonal cells) sequentially using span features computed from contextualized representations and then predicts the relation labels (the off-diagonal cells) simultaneously using the scores of each word pair computed by the tensor dot-product (Ma et al., 2022).

We trained two models to extract the part entity and relations and to extract causality. Subsequently, we used part relations as external knowledge when training the table-filling model to verify how the entity and relation knowledge help improve the causality extraction. Employing the PURE (Zhong & Chen, 2021) method, we can add the information of part entities and relations to the end of the input text without modifying the architecture of the table-filling model. Specifically, we added a sequence containing all entity pairs with relations. For each entity pair, we added markers to the end of the text:

<E:S></E:S>relation_type<E:O></E:O>.

Here, <E:S> and </E:S> share the same position embedding with the start and end tokens of the subject (head) entity. Similarly, <E:O> and </E:O> specify the start and end positions of the object (tail) entity, respectively.

4.2 Experimental Settings

The annotated data were randomly divided into training and evaluation data in an 8:2 ratio. The open-source implementation of the table-filling model[5] was used in the experiments. We employed Japanese Bert as a pre-trained model[6]. The experiments were conducted on a single GPU of NVIDIA GTX 1080 Ti (11 GiB.) For hyperparameter settings, we set the learning rate to 5×10^{-5}, the dropout rate to 0.3, and the maximum length of the input tokens to 250.

4.3 Evaluation results

First, we present the experimental results for NER and RE tasks on the corpus annotated with NEs and relations in Table 3. Precision (P), recall (R), and F1-score (F1) were used for the evaluation. For part entity extraction, the model reached a high F1 score of 0.9710, and the precision and recall were also as high as the F1-score. The high F1-score indicates that the model can recognize the vehicle entity spans. Next, we observe the best overall performance for COREFERENCE and the lowest accuracy for ONEWAY-COREFERENCE. Their performance reflects the number of their occurrences in the corpus as the type COREFERENCE is the most frequent and type ONEWAY-COREFERENCE is the least frequent in the corpus (Table 1). Notably, although the number of relation instances annotated as COREFERENCE and PART-WHOLE are similar, the F1-score for PART-WHOLE is approximately 0.19 less than that for COREFERENCE. Moreover, the model does not perform well in predicting CONNECT and CONTACT, with f1-score 0.5045 and 0.6371, respectively.

Type	P	R	F1
NER			
PARTS	0.9674	0.9746	0.9710
RE			
CONNECT	0.5636	0.4566	0.5045
COREFERENCE	0.8972	0.9230	0.9099
ONEWAY-COREFERENCE	0.6042	0.3187	0.4173
PART-WHOLE	0.7221	0.7018	0.7118
CONTACT	0.6630	0.6131	0.6371
All	0.7585	0.7270	0.7424

Table 3: Experimental results on NER and RE.

	P	R	F1
TablERT	0.7120	0.7341	0.7229
TablERT + parts_info	**0.7193**	**0.7389**	**0.7290**

Table 4: Experimental results on causality extraction

Next, we present the experimental results for the causality extraction task in Table 4. The first row summarizes the results of causality extraction without entity and relation knowledge, and the second row summarizes the results of causality extraction with the external entity and relation information. Compared to the model without any external knowledge, including the part entity and relation knowledge improves all evaluation metrics. Although the

[5] https://github.com/YoumiMa/Enhanced_TF

[6] https://github.com/cl-tohoku/bert-japanese

improvement is modest, the results show that entities and relations can enhance the performance of causality extraction on the vehicle recall corpus.

4.4 Discussion

It is challenging for the model to extract the implicit relations. Although the model performs well on the NER task on the part-relation corpus, there is still room for improvement regarding RE, especially in predicting certain relation types.

For type COREFERENCE, the model reached an impressive f1-score of 0.9099. One possible reason for the high performance is the property that two PARTS entities paired with a COREFERENCE relation usually share a common word span. As presented in Table 3, the model can recognize word spans with high accuracy, resulting in a high accuracy in extracting COREFERENCE tuples as these tasks are similar. For instance, in Figure 1, PARTS "auxiliary braking device (electromagnetic retarder)" and "electromagnetic retarder" have a COREFERENCE relation and they share the common word span "electromagnetic retarder."

In contrast, the F1 scores for the other relation types were relatively low, possibly because knowledge about these relation types is usually absent from the document. Again, take the annotated document illustrated in Figure 1 as an example, "electromagnetic retarder" and "braking light" has a CONNECT relation; however, recognizing the relation is difficult even for a non-expert human. Thus, it is understandable that the problem is also difficult for a machine learning system without access to expertise in vehicles.

5 Conclusion

In this work, we have presented two annotations on the vehicle recall dataset: NE and relation annotation and causality annotation. We trained NER and RE models for the NER and RE and causality extraction tasks on both annotated corpora and presented the results. The results demonstrate the feasibility of building a causality relation system using an annotated corpus. Subsequently, we used the part entity and relation annotated corpus to improve causality extraction on the car recall corpus. The experimental results show that incorporating part entity knowledge improves the performance of causality extraction.

In future work, we will investigate a more effective approach to utilize NE and relation information to improve causality extraction.

Acknowledgments

This paper is based on the results obtained from a project, JPNP18002, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating Causal Language Using Corpus Lexicography of Constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2, In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.*

2002. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain.*, Association for Computational Linguistics, Phildadelphia, Pennsylvania, USA, edition.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support?, *Journal of the Association for Computing Machinery.*

Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann. 2016. *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, The COLING 2016 Organizing Committee, Osaka, Japan, edition.

2020. Natural language processing applications in business. Mohammed Bahja, ed. by *E-Business-Higher Education and Intelligence Applications*, IntechOpen.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *EACL,* pages 102–107.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007),* pages 13–18.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco

Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation,* pages 33–38.

Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. Web Download. Philadelphia: Linguistic Data Consortium, 2008.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In *Proceedings of the 11th Linguistic Annotation Workshop,* pages 95–104.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events,* pages 51–61.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 839–849.

James Pustejovsky, José Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering,* page 28-34.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating Causality in the TempEval-3 Corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL),* pages 10–19.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics,* 8(1), pages 1-24.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics,* 14(1), pages 1-18.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania.*

Youmi Ma, Tatsuya Hiraoka,and Naoaki Okazaki. 2020. Named Entity Recognition and Relation Extraction Using Enhanced Table Filling by Contextualized Representations. *Journal of Natural Language Processing,* 29(1), pages 187–223.

Dan Roth and Wen-tau Yih. 2004. Annotating Causality in the TempEval-3 Corpus. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004,* pages 1–8.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling Joint Entity and Relation Extraction with Table Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pages 1858–1869.

Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009),* pages 147–155.

Zexuan Zhong,and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 50–61.

A Deep Learning based Framework for Image Paragraph Generation in Hindi

Santosh Kumar Mishra[1], **Sushant Sinha** [1], **Sriparna Saha**[1], **Pushpak Bhattacharyya**[2]
Indian Institute of Technnology Patna [1] , Indian Institute of Technnology Bombay [2]
{santosh_1821cs03, sushant_1901cs62, sriparna}@iitp.ac.in[1],pb@cse.iitb.ac.in[2]

Abstract

Generating a paragraph from an image is a complex task that requires object and activity recognition. It is now feasible because of recent advances in image captioning. Summarizing an image into a single sentence can only provide a superficial description of the visual information included within the image. This problem can be solved by generating a detailed and coherent description of the input image. Most existing works on image-to-paragraph generation have been accomplished in English. We propose a novel way of generating a paragraph from an image in Hindi. The world's third most spoken language and one of India's official languages, Hindi, is extensively spoken throughout India and South Asia. We construct a new dataset for an image-to-paragraph generation in Hindi. We employ a hierarchical recurrent neural network (HRNN) for language modeling and an object detection model to decompose the input image by distinguishing regions of interest and objects. The performance of the proposed methodology is compared with other baselines in terms of BLEU, CIDER, and METEOR scores, and the obtained results show that the proposed method outperforms them.

1 Introduction

Human perception is dependent on vision and language. These are the most effective means of interacting with others. In our everyday lives, we are exposed to a large number of images from various sources such as advertisements, news stories, and the internet, among others. As a result, a technique to define the visual information and activity inside an image is essential. A major challenge in computer vision and natural language processing is developing a system that can explain visual objects and their relationships in natural language. A computer-generated image description may aid visually impaired people in comprehending online information (MacLeod et al., 2017). Recent advances in artificial intelligence, computer vision, and natural language processing have enabled image or video captioning to explain the visual content of an image or video (Vinyals et al., 2015) (Vinyals et al., 2016) (Anderson et al., 2018), (Anderson et al., 2018) Because of the availability of large datasets (Young et al., 2014) (Lin et al., 2014) that combine images with natural language descriptions, it is now possible to generate sentences to interpret the images. Though the effectiveness of these solutions is promising, they all have one big flaw: they all fail to capture the subtleties inside an image or video.

The generated caption in the image or video captioning procedure is a single sentence of around 20 words that anticipates just the essential observations within an image. Capturing all of the visual information and semantics included inside an image or video is inadequate. This problem can be solved by generating a paragraph description from an image that has comprehensive visual information. When compared to single-sentence captioning, the generation of a paragraph from images is a relatively new research area. Visual Genome, an image-to-paragraph generation dataset introduced by (Krause et al., 2017) plays an important role in the image-to-paragraph generation. This data set can be used for training any machine learning classifier to develop an image captioning model. But, machine learning models can not generate a precise paragraph for a variety of images. They generate repetitive sentences very often. To resolve this issue, most researchers have used hierarchical LSTMs, which generate separate words and sentence topics. Recent research works on image-to-paragraph generation (Johnson et al., 2016)(Krause et al., 2017)(Yu et al., 2016) (Liang et al., 2017) provide a bigger narrative while generating the description for an image or video. Image-to-paragraph generation is a challenging task that requires understanding images and language modeling.

Previously, image-to-paragraph generation work was limited to the English language. In this paper, we present the first image-to-paragraph model in the Hindi language. As encoder and decoder, we use Faster R-CNN and hierarchical recurrent neural network, respectively. This work has made the following significant contributions:

- This is the first effort of its kind for paragraph generation from an image in Hindi. We use Faster R-CNN as an encoder to decompose the input image by recognizing distinct objects and regions of interest. Furthermore, the features of these regions are combined to form a rich representation of image semantics. As a decoder, a hierarchical neural network composed of sentence RNN and word RNN is employed; it uses these rich representations for language modeling.
- We present a novel Hindi dataset for paragraph generation from an image by translating a well-known Visual Genome dataset (Krause et al., 2017). Using Google Translate, we translate the whole corpus from English to Hindi. Furthermore, human annotators correct the translation according to Hindi grammatical rules, which requires a substantial amount of human effort and time.

2 Related Works

Many studies have been carried out in the past to combine visual and textual data. It has been accomplished in a variety of methods throughout the literature. Some researchers have addressed this as a ranking issue, using the image as input to identify the appropriate caption from the dataset and vice versa (Farhadi et al., 2010) (Hodosh et al., 2013).

An encoder-decoder architecture is used in nearly all recent image captioning models in the literature. The encoder is a CNN architecture pre-trained for image classification, while the decoder is mainly an LSTM or GRU as proposed by (Vinyals et al., 2015). In most cases, a convolutional neural network (CNN) is used to generate an encoding of the given source image. After that, the image encoding is put into an RNN, which selects a collection of words (from a dictionary) that match the most with the image encoding. In (Xu et al., 2015), authors have employed RNNs as a decoder with an attention mechanism for caption generation from images; this mechanism focuses on the relevant parts of the image while generating the caption. Using a faster R-CNN (Ren et al., 2015) object detection model, in (Anderson et al., 2018), bottom-up and top-down attention mechanisms are introduced for caption generation from images. A modified encoder-decoder model with a guiding network is also utilized for image captioning (Jiang et al., 2018), the data to the decoder at each time step is the output of the guiding network. An unsupervised method of learning for image caption generation is introduced in (Feng et al., 2019), the proposed model did not employ image and sentence pairs for image captioning. A meshed memory transformer network is introduced for image captioning in (Cornia et al., 2020), it uses a multi-level representation of the region's relationship with prior information. An image captioning model based on ensemble generation and retrieval using generative adversarial networks is explored in (Liu et al., 2020). A language pre-training model's unified version is developed in (Zhou et al., 2020); it accomplishes language modeling based on the shared transformer network. The captions produced by the above methods are generally brief, comprising only a single phrase of no more than 20 words.

Intuitively, image to paragraph generation appears to be similar to image captioning: given an image, generate a written description of its content (Krause et al., 2017). The inventiveness in the textual description, on the other hand, is essential for the image to paragraph generation. The image-to-paragraph generation framework, in particular, is intended to generate a paragraph consisting of five or six sentences that describe the image in more detail. Furthermore, a seamless transition between the sentences of the paragraph's phrases is required. Authors of (Johnson et al., 2016) proposed a method for producing comprehensive captions. A focus on a story theme underlying a specific image was lacking while producing engaging words separately. In [21], the authors proposed a method to deal with this issue. A two-stage hierarchy of RNNs is used in their language model. Given a visual representation of semantically significant areas in an image, the first

RNN level generates a sentence vector. This subject vector is converted into a sentence at the second RNN level. They released the first large-scale image-to-paragraph generation dataset, a subset of the Visual Genome dataset, as well as many paragraph captioning algorithms. The author of (Liang et al., 2017) added a third (paragraph-level) LSTM to this model (Krause et al., 2017), as well as adversarial training. Three LSTMs, two attention mechanisms, a phrase copy mechanism, and two adversarial discriminators are all included in their model (RTT-GAN).

Previously, the majority of studies were undertaken simply for the generation of paragraphs from images in English. To the best of our knowledge, no attempt has been made to generate paragraphs from images in Hindi. Our methodology is the first of its kind that generates paragraphs from images in Hindi.

3 Proposed Method

The proposed method takes an image as input and generates a natural language paragraph description of the image, making use of the compositional structure of both images and paragraphs (as illustrated in Fig 1). It deconstructs the input image by recognizing objects inside and other regions of interest and then combines features from all of these components to construct a pooled representation that reflects the image semantics.

A hierarchical recurrent neural network comprising two levels: a sentence RNN and a word RNN, takes this feature vector as input. The image features are sent to the sentence RNN, which then determines how many sentences to generate in the resulting paragraph and generates an input topic vector for each sentence. The word RNN generates the words of a single sentence given this topic vector. This section has a brief explanation of each of these modules, which are as follows:

3.1 Detection of Regions using Region Proposal Network

The proposed method uses a region proposal network (RPN) to detect regions of interest (ROI) as introduced in (Ren et al., 2015). It takes an input image of dimension $3 \times H \times W$ and finds regions of interest, and generates a D = 4096 feature vector for each region. H and W are the height and width of the image, respectively. A convolutional neural network using the VGG-16 network processes the input image. It generates a feature map, which is subsequently processed by a region proposal network that regresses from a group of anchors. The region detector is trained in an end-to-end manner (Ren et al., 2015) for object recognition and for dense image captioning as well (Johnson et al., 2016), given a dataset consisting of captions with areas of interest. The region detector is trained for object detection (Ren et al., 2015) is utilized for the dense image captioning model (Johnson et al., 2016), using a dataset of images and corresponding ground ROI. We employ a region detector trained for dense caption generation of images on the visual genome dataset (Krishna et al., 2016), utilizing a publicly available implementation of (Krause et al., 2017) because the paragraph description does not contain annotated grounding to ROI (region of interest).

3.2 Region Pooling

The region proposal network detects different regions and generates a set of vectors $v_1, \ldots, v_M \in R^D$, denoting various regions in the input images. These vectors are aggregated into a pooled vector $v_p \in R^p$, which describes the content of an image. Pooled vector v_p is computed using element-wise maximum as follows:

$$v_p = max_{i=1}^{M}(W_{pool}v_i + b_{pool}) \tag{1}$$

Here $W_{pool} \in R^{P \times D}$ is a learned projection matrix, and bias $b_{pool} \in R^P$ is the bias.

3.3 Language Modeling Hierarchical Recurrent Neural Network

An HRNN based language model consists of two components: word and sentence RNN. The number of sentences to be generated is decided by the sentence RNN; it generates a topic vector of dimension P. A hierarchical neural language model is given the pooled region vector $v_p \in R^P$ as input. The word RNN produces words for a sentence given a topic vector. For both word RNN and sentence RNN, we use the conventional LSTM architecture (Hochreiter and Schmidhuber, 1997).

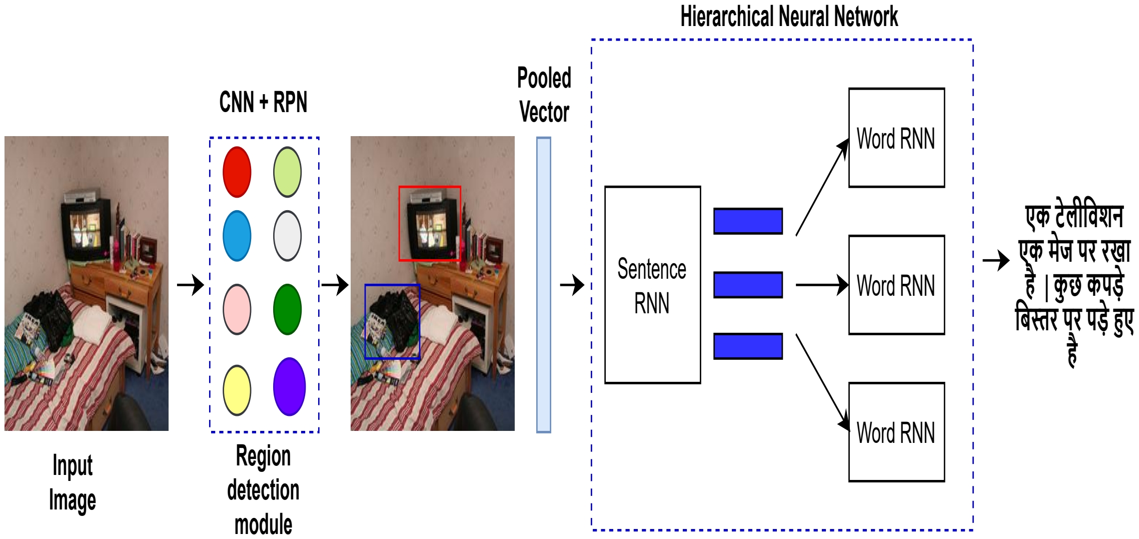


Figure 1: Architecture of the proposed method

Sentence RNN is an LSTM consisting of a single layer; its cells and hidden states are initialized with zero with a hidden size $H = 512$. At each time step, the pooling region vector v_p is given as an input to the sentence RNN, and it produces a series of hidden states, H, one for each sentence in the paragraph. The word RNN consists of two-layered LSTM with a hidden size of $H = 512$; it generates the words of the sentence given the topic vector, $t_i \in R^p$. The subject vector and a special START token are the RNN's initial and second inputs, respectively, while successive inputs are learned embedding vectors for the sentence's words. A unique END token indicates the end of a sentence, and the hidden state of the final LSTM layer is utilized to forecast a distribution across the words in the vocabulary at each timestep. Following the generation of the words for each sentence by word RNN, the sentences are combined to form the resulting paragraph.

3.4 Training Procedure

This section provides a detailed description of the training procedure. The training data is made up of pairs (x, y) where x and y represent an image and a corresponding ground truth paragraph description, respectively. Here, y consists of S number of sentences. The i^{th} sentence has N_i words and y_{ij} denotes the j^{th} word of the i^{th} sentence. Sentence RNN is unrolled for S sentences after a pooled region vector v_p is computed for an image. For each sentence, the sentence RNN generates the probability distribution p_i over the $CONTINUE, STOP$. Here, $CONTINUE, STOP$ are the special keywords to determine when to stop or continue generating the sentences in the paragraph.

Training loss $L(X, Y)$ is the weighted summation of word loss L_{word} and sentence loss $L_{sentence}$. It is defined as follows:

$$\begin{aligned} L(x, y) = \lambda_{sent} \sum_{i=1}^{S} L_{sent}(p_i, I[i = S]) \\ + \lambda_{word} \sum_{i=1}^{S} \sum_{j=1}^{N_i} L_{word}(p_{ij}, y_{ij}) \end{aligned} \tag{2}$$

Here, the sentence RNN generates the sentences until it reaches S_{max} or stopping probability, $p_i(STOP)$, exceeds a threshold, T_{stop}. Here, values of the above parameters are as follows; T_{stop} = 0.5, $S_{Max} = 6$ and $N_{MAX} = 50$. We also incorporate self-critical sequence training in the above architecture to enhance the diversity in the paragraph generation (Rennie et al., 2017).

4 Experimental Setup

4.1 Dataset

We construct a dataset for the task of paragraph generation from images in Hindi by translating the well-known Stanford image to paragraph generation dataset (Krause et al., 2017) from English to Hindi, [1]. It has a total of 19,551 images taken from the MSCOCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2016) data sets. The dataset is divided into 3 parts: 14,575 training images, 2,487 validation images, and 2,489 testing images. Initially, all of the English captions were translated into Hindi by Google Translator; the following issues were encountered while translating from English to Hindi:

- Since Google Translate lacks a technique for adding sentence context, the translated caption's meaning is lost during translation.
- In certain cases, Google Translator's translation is grammatically incorrect.
- The Google Translator accuracy is not uniform as it is dependent on pairings of source and target languages.

As a result, human annotators are employed to correct Google-translated sentences to remove errors. The inter-annotator agreement was 87% between the two annotators. A sample example from the dataset is shown in Fig 2 and Fig 3.

Though we can get the paragraph for an image in Hindi by translating the English paragraph generated by a model trained for English, the resultant caption lacks adequacy and fluency, as shown by the authors of (Mishra et al., 2021a) (Mishra et al., 2021b). This demonstrates the need of constructing a Hindi dataset for image-to-paragraph generation.

4.2 Evaluation Metrics

We evaluate the proposed methodology using BLEU (Papineni et al., 2002), CIDEr-D (Vedantam et al., 2015), and METEOR (Denkowski and Lavie, 2014).

4.3 Hyper-parameters Used

The proposed architecture incorporates two layers of LSTM with a dimension of 512. The dimension of feature pooling is 1024. Stochastic gradient descent with the Adam optimizer (Kingma and Ba, 2014) is used for training. Values of λ_{sent} and λ_{word} are set to 5.0 and 1.0, respectively. The model has been trained for 30 epochs, which takes approximately 12 hours of training.

5 Results and Discussion

This section covers the detailed discussions of the results and analysis. We carried out the experimentation on the introduced image-to-paragraph generation dataset in Hindi.

5.1 Comparative baselines for image to paragraph generation

To the best of our understanding, no work has been done on paragraph generation from images in the Hindi language. Therefore, we create our own baselines, which are as follows:

- **Baseline -1:** In this baseline, top-down attention (Anderson et al., 2018) is incorporated with Faster-R CNN (Ren et al., 2015) and bi-LSTM (Hochreiter and Schmidhuber, 1997), here we explore the bi-directional LSTM for language modeling.
- **Baseline-2:** In this baseline, adaptive attention (Lu et al., 2017) is incorporated with Faster-R CNN (Ren et al., 2015) and LSTM (Hochreiter and Schmidhuber, 1997).
- **Baseline-3:** In this, we explore adaptive attention (Lu et al., 2017) with Faster-R CNN (Ren et al., 2015) and LSTM (Hochreiter and Schmidhuber, 1997) with Maxout (MO) activation function (Goodfellow et al., 2013).

5.2 Qualitative Evaluation

This section shows a qualitative evaluation of the proposed methodology on test images. The generated paragraph for the test image is shown in Fig 4. We include the transliteration and gloss annotation so that non-Hindi speakers can grasp the meanings of the captions. It can be seen from the Fig 4 that the

[1] dataset will be released on acceptance

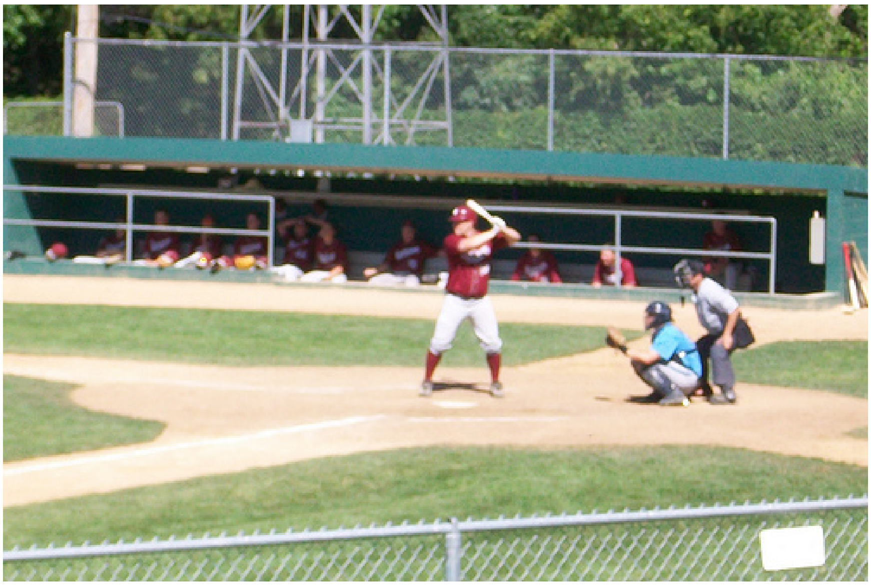
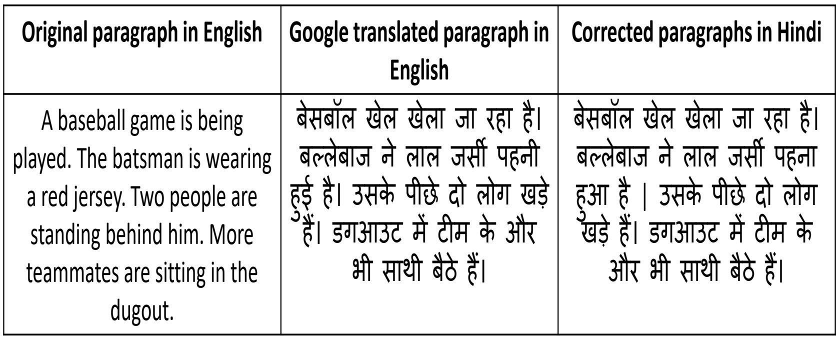


Figure 2: A sample image from the dataset created in Hindi



Original paragraph in English	Google translated paragraph in English	Corrected paragraphs in Hindi
A baseball game is being played. The batsman is wearing a red jersey. Two people are standing behind him. More teammates are sitting in the dugout.	बेसबॉल खेल खेला जा रहा है। बल्लेबाज ने लाल जर्सी पहनी हुई है। उसके पीछे दो लोग खड़े हैं। डगआउट में टीम के और भी साथी बैठे हैं।	बेसबॉल खेल खेला जा रहा है। बल्लेबाज ने लाल जर्सी पहना हुआ है | उसके पीछे दो लोग खड़े हैं। डगआउट में टीम के और भी साथी बैठे हैं।

Figure 3: A sample paragraph for the given image from the dataset created in Hindi

State-of-the-art/baselines	Language modeling	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
Top Down Attention (Proposed Method)	LSTM	18.783	6.766	11.688	19.974	35.792	27.490	26.71
Top Down Attention (Baseline-1)	Bi- LSTM	12.666	5.220	9.471	16.930	31.063	27.545	23.684
Adaptive Attention (Baseline-2)	LSTM	17.111	5.838	10.648	19.017	34.948	27.292	26.92
Adaptive Attention MO (Baseline-3)	LSTM	20.74	5.954	10.895	19.339	34.91	27.149	26.426

Table 1: Obtained score with proposed method and baselines

State-of-the-arts/Baselines	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
Baseline-1	2.24261e-109	6.40514e-86	1.21712e-86	7.78625e-98	5.22064e-105	4.00075e-68	1.98766e-102
Baseline -2	5.69975e-89	1.07877e-77	1.9483e-79	1.66527e-91	3.27974e-88	1.00585e-52	5.16477e-81
Baseline-3	6.14475e-91	9.60087e-76	3.29491e-87	1.60837e-79	2.40272e-102	1.9483e-79	1.383e-111

Table 2: Welch's t-test based comparison between proposed method and state-of-the-art baselines.

produced captions for the test photographs are pretty accurate and appropriately identify the actions and items in the images.

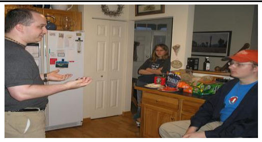
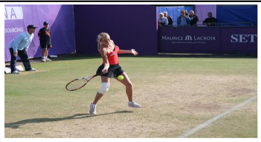
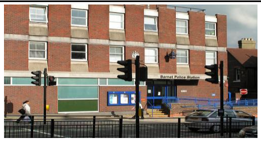
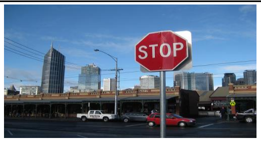
			
(a) I - एक आदमी एक रसोई में एक रसोई के अंदर खड़ा है आदमी के सामने एक सफेद है । आदमी के पास एक सफेद शर्ट है । महिला के पास एक सफेद शर्ट है । आदमी एक सफेद रंग की शर्ट पहने हुए है । वह एक सफेद रेफ्रिजरेटर के सामने खड़ा है । एक सफेद दीवार के सामने बैठा है । कमरे के सामने एक दीवार है । दीवार पर एक सफेद है । II- One man one kitchen in one kitchen standing inside man of in front white. Man of has one white shirt. Women of one white shirt. Man a white shirt wearing. He one standing in front of white refrigerator. He white wall in front of seating. Room in front one wall is. Wall of one white is. III- Ek adami ek rasoi me ek rasoi ke andar khara hai adami ke samne ek safed hai. Adami ke pas ek safed shirt hai. Mahila ke pas ek safed shirt hai . Adami ek safed rang ki shirt pahne hue hai. Vah ek safed refrigerator ke samne khara hai. Vah safed deewar ke samne baitha hai. Kamre ke samne ek deewar hai. Deewar par ek safed hai.	(b) I - एक महिला टेनिस कोर्ट पर खड़ी है । उसने एक सफेद शर्ट और सफेद शॉर्ट्स पहने हुए है । महिला ने एक सफेद टैंक टॉप और सफेद स्कर्ट पहन रखी है । वह एक सफेद रैकेट पकड़े हुए है । महिला के पीछे एक सफेद है । लड़की के पीछे एक बड़ा सफेद है । अदालत के पीछे एक आदमी खड़ा है । एक सफेद रंग की टेनिस कोर्ट है । II- One women tennis court on standing. She one white shirt and white shorts wearing. Women of one white top and white skirt wearing. Women of behind one white. Girl behind one big white is. Court of behind one man standing. One white color of tennis court is. III- Ek mahila tennis court par khari hai. Usne ek safed shirt and safed shorts wearing. Mahila ne ek safed tak top aur safed racket pakre hue hai. Mahila ke piche ek safed hai. Ladaki ke piche ek bara safed hai. Adalat ke piche ek adami khara hai. Ek safed rang ki tennis court hai.	(C) I- एक इमारत के सामने एक सड़क है । भवन के सामने एक पोल है । इमारत के सामने सड़क पर एक पोल है । सड़क के किनारे एक पोल पर एक सफेद है । गली के सामने एक सफेद पोल है । भवन के सामने सड़क के एक सफेद कार है । एक के सामने फुटपाथ पर एक काला पोल है है । पोल के सामने एक इमारत है । II- One building of in front of one road is. Building in front of one poll is. Building in front of road on one poll is. Road of side one poll on one white is . Lane in front of one white poll is. Building in front of road on one white car is. One in front of footpath on one black poll is. Poll in front of one building is. III – Ek imarat ke samne ek sadak hai. Bhawan ke samne ek poll hai. Imarat ke samne sadak par ek poll hai. Sadak ke kinare ek poll par ek safed hai. Gali ke kinare ek safed poll hai. Bhawan ke samne sadak ke ek safed kar hai. Ek ke samne footpath par ek kala poll hai. Poll ke samne ek imarat hai.	(d) I - एक सड़क के किनारे एक स्टॉप साइन है । स्टॉप साइन के सामने एक सफेद है । संकेत के सामने एक स्टॉप साइन है । सड़क के बगल में एक सफेद ट्रक है । सड़क के सामने सड़क पर एक सफेद कार है । गली के सामने सड़क के एक सफेद वैन है । इमारत के सामने एक सड़क है । एक के पीछे एक सफेद इमारत है । II- One road of side one stop sign is. Stop sign in front of white is. Indication in front of one stop sign is. Beside road one white truck is. In front of road on road one white van is. Building in front of one white is. One of behind one white building is. III- Ek sadak ke kinare ek stop sign hai. Stop sign ke samne ek safed hai. Sanket ke samne ek stop sign hai. Sadak ke bagal me ek safed truck hai. Sadak ke samne sadak par ek safed car hai. Gali ke samne sadak ke ek safed van hai. Imarat ke samne ek sadak hai. Ek ke piche ek safed imarat hai.

Figure 4: Generated paragraph by the proposed method on test images. Here, I, II and III denote the Hindi generated caption, gloss annotation and transliteration, respectively.

5.3 Quantitative Analysis

Although the qualitative analysis has been carried out manually, to conduct the quantitative analysis, a subjective score is still required. The generated Hindi paragraphs here are evaluated against the ground truth paragraph. We perform the qualitative analysis using the BLEU score (Papineni et al., 2002); using n-grams, METEOR (Denkowski and Lavie, 2014), and CIDEr (Vedantam et al., 2015) scores.

We validate our proposed approach and compared it to different baselines using BLEU, CIDEr, and METEOR scores, as shown in Table 1. The results show that our proposed approach outperforms all current baselines.

5.4 Statistical Significance Test

We conduct a statistical significance test (Welch, 1947) at a 5% (0.05) significance level to ensure that the performance increase achieved by our technique is statistically significant. This test provides the p-values; the lower the p-values, the greater the significance compared to state-of-the-art approaches. We obtain all of the values less than 0.05 (as shown in Table 2), establishing the statistical significance of our technique and demonstrating that the improvement gained by the proposed technique is not by coincidence.

6 Conclusion and Future Works

We present a novel framework for generating paragraphs from photographs in Hindi, which incorporates a region proposal network-based convolutional neural network and an LSTM-based encoder-decoder model with attention mechanisms. We analyze various encoder-decoder models to find the best architecture for paragraph generation from images in Hindi. This work could be extended further by using a transformer-based architecture for language modeling.

Acknowledgment Dr. Sriparna Saha would like to acknowledge the support received from the Young Faculty Research Fellowship program of Visvesvaraya Ph.D. Scheme of Ministry of Electronics Information Technology, Government of India, being implemented by Digital India Corporation (Formerly Media Lab Asia) for conducting this research.

References

[Anderson et al.2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

[Cornia et al.2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

[Denkowski and Lavie2014] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

[Farhadi et al.2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

[Feng et al.2019] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

[Goodfellow et al.2013] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hodosh et al.2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

[Jiang et al.2018] Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018. Learning to guide decoding for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Johnson et al.2016] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

[Kingma and Ba2014] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Krause et al.2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

[Krishna et al.2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

[Liang et al.2017] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371.

[Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[Liu et al.2020] Junhao Liu, Kai Wang, Chunpu Xu, Zhou Zhao, Ruifeng Xu, Ying Shen, and Min Yang. 2020. Interactive dual generative adversarial networks for image captioning. In *AAAI*, pages 11588–11595.

[Lu et al.2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

[MacLeod et al.2017] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999. ACM.

[Mishra et al.2021a] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, and Pushpak Bhattacharyya. 2021a. A hindi image caption generation framework using deep learning. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–19.

[Mishra et al.2021b] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, and Amit Kumar Singh. 2021b. Image captioning in hindi language using transformer networks. *Computers & Electrical Engineering*, 92:107114.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Ren et al.2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

[Rennie et al.2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

[Vedantam et al.2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

[Vinyals et al.2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[Vinyals et al.2016] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

[Welch1947] Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

[Xu et al.2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

[Young et al.2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

[Yu et al.2016] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

[Zhou et al.2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049.

UIT-ViCoV19QA: A Dataset for COVID-19 Community-based Question Answering on Vietnamese Language

Triet Minh Thai
University of Information Technology
VNU-HCM, Vietnam
19522397@gm.uit.edu.vn

Ngan Ha-Thao Chu
University of Information Technology
VNU-HCM, Vietnam
19521882@gm.uit.edu.vn

Anh Tuan Vo
University of Information Technology
VNU-HCM, Vietnam
19521226@gm.uit.edu.vn

Son T. Luu
University of Information Technology
VNU-HCM, Vietnam
sonlt@uit.edu.vn

Abstract

For the last two years, from 2020 to 2021, COVID-19 has broken disease prevention measures in many countries, including Vietnam, and negatively impacted various aspects of human life and the social community. Besides, the misleading information in the community and fake news about the pandemic are also serious situations. Therefore, we present the first Vietnamese community-based question answering dataset for developing question answering systems for COVID-19 called UIT-ViCoV19QA. The dataset comprises 4,500 question-answer pairs collected from trusted medical sources, with at least one answer and at most four unique paraphrased answers per question. Along with the dataset, we set up various deep learning models as baseline to assess the quality of our dataset and initiate the benchmark results for further research through commonly used metrics such as BLEU, METEOR, and ROUGE-L. We also illustrate the positive effects of having multiple paraphrased answers experimented on these models, especially on Transformer - a dominant architecture in the field of study.

1 Introduction

Community-based Question answering (CQA) is a task of question answering based on the wisdom of the crowd (Zhang et al., 2021). In CQA, information seekers post their questions on a public website or forum, and other users answer them. This kind of question-answering behavior is popular in people's daily basis. For example, Quora[1] and Reddit[2] are several large forums for question-answering. The CQA enables people to ask and answer questions easily (Qiu and Huang, 2015). Additionally, the development of question and answering systems means computers can now understand and answer the questions of users.

In Vietnam, the year from 2020 to 2021 witnessed the COVID-19 pandemic. Information about the COVID-19 spreading situation, medical care, self-quarantine, vaccination policies, and regulations by the government to prevent the spread of COVID-19 are essential to citizens. People frequently ask questions about the COVID-19 situation, what to do when contacting COVID-19 patients, the vaccination policies, and more. This is our motivation to construct a dataset to help build a question answering system based on CQA about COVID-19 in Vietnamese. Apart from the dataset, we also propose various baseline models to evaluate our dataset's quality. This paper has three main contributions summarized as follows:

1. We introduce UIT-ViCoV19QA, the first community-based question answering collection about the COVID-19 pandemic for Vietnamese constructed from trusted sources. The dataset comprises 4,500 question-answer pairs and is extended to have up to four unique paraphrased answers per question through an efficient paraphrasing process.

2. We assess the dataset's quality and estab-

[1] https://www.quora.com/
[2] https://www.reddit.com/

lish a future research benchmark through experiments with various Sequence-to-Sequence baselines to automatically generate answer for a given question about COVID-19 in Vietnamese.

3. We perform error analysis and illustrate that models trained on multiple paraphrased answers tend to have better generalization than those trained using only one answer. This reflects the advantage of having multiple paraphrased answers in single-turn conversational question answering.

The rest of this paper is structured as follows. In the following section, we review the related works. In Section 3, we describe the process of building UIT-ViCoV19QA dataset in detail, including data collection, data pre-processing, paraphrasing process, and highlight its overall statistics. Section 4 is devoted to methodologies and experiment configurations. The results and benchmarks, as well as error analysis are described in Section 5. Finally, our conclusion and future works are presented in Section 6.

2 Related works

Question answering systems consist of the single-turn and multi-turn QA. According to (Del Tredici et al., 2021), the single-turn QA takes the questions as input and returns the output without context. Single-turn QA includes Text-based QA (SQuAD (Rajpurkar et al., 2016), RACE (Lai et al., 2017)), Visual-QA (VQA (Goyal et al., 2017)), Community-based QA (ANTIQUE (Hashemi et al., 2020)), and Knowledge-based QA (MetaQA (Zhang et al., 2018)) In contrast, the multi-turn QA take the questions as input belong with contexts such as conversation history, which is called the Conversational QA (CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018)). Additionally, to verbalize the response from the question, the ParaQA dataset (Kacupaj et al., 2021) used the paraphrasing techniques. Each question in the ParaQA dataset has at least two more answers. In particular, some questions contain eight responses.

On the other hand, many efforts to create the question answering corpora in Vietnamese such as UIT-ViQuAD (Nguyen et al., 2020), UIT-ViNewsQA (Van Nguyen et al., 2020b), ViMMRC (Van Nguyen et al., 2020a) for text-based question answering, ViVQA (Tran et al., 2021) for visual-based QA, and ViCoQA (Luu et al., 2021) for conversational QA. Our works contributes to the corpora for Vietnamese as Community-based QA dataset. With the idea from the ParaQA dataset (Kacupaj et al., 2021), we manually create up to three responses from initial answer for each question in our dataset to make the dataset verbalized.

3 Dataset

In order to publish a high-quality dataset for the research community and be able to experiment with the baseline models, we have investigated building a dataset for Vietnamese, a low-resource language in the field of study. This section describes the construction of UIT-ViCoV19QA in detail and presents some statistics from the dataset. The overview of the dataset construction workflow is illustrated in Figure 1.

3.1 Data Collection

Question-answer pairs used for the constructing of UIT-ViCoV19QA dataset are extracted from FAQ documents that are publicly available on websites of respected health care organizations in Vietnam and overseas, including The Centers for Disease Control and Prevention (CDC), United Nations Children's Fund (UNICEF), The Ministry of Health of Vietnam, Vietnam Government Portal and other trusted medical institutions. Each web page provides different topics about the COVID-19 pandemic compiled in Vietnamese and often includes similar information extracted during the crawl. The following topics are covered in our dataset: origin, outbreak, and name of the disease; spread; symptoms; prevention, treatment guidelines, and nutrition; treatment models; variants of COVID-19; vaccines and vaccination; moving between areas, entry, and travel; isolation, quarantine, lockdown and social distancing; policies and sanctions; financial support; post-COVID-19; COVID-19 in children. Disease statistics in Vietnam and worldwide are not included in the dataset.

Once appropriate and trustful sources are identified, numerous handcrafted patterns are developed

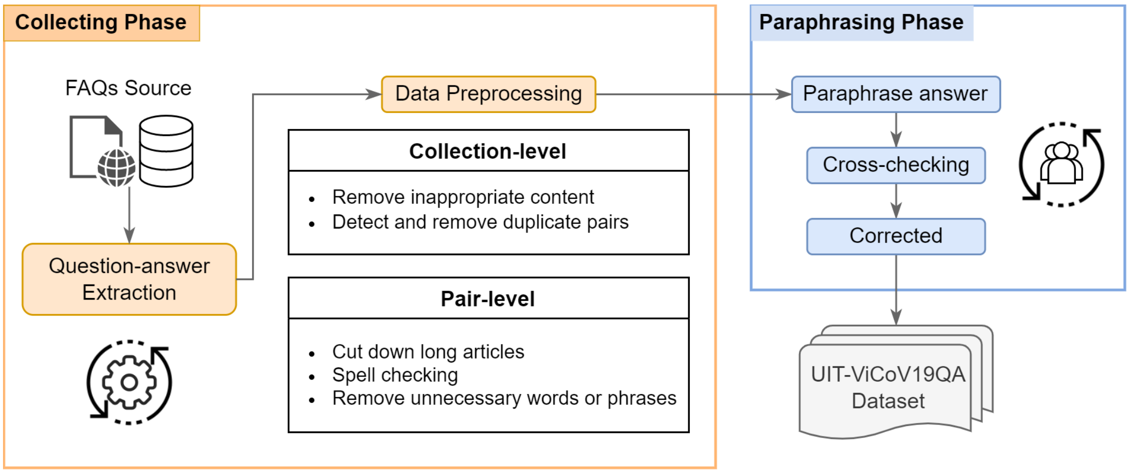


Figure 1: An overview of UIT-ViCoV19QA constructing workflow

for each website to automatically extract question-answer pairs based on the document structure, typically in HTML and PDF format. Throughout the collecting phase, the length of both question and answers are maintained to retain information. Therefore, the collected content could be a sentence, a paragraph, or a passage determined by the author or compiler. This poses a tremendous challenge for Sequence-to-Sequence models to produce good results, which makes it an ideal material for further investigations.

3.2 Data Pre-processing

A two-stage pre-processing is performed on the collected question-answer pairs. First, duplicate pairs are detected using Cosine similarity at the collection-level and are considered to be taken out of the collection. During the crawl, we also attempt to identify and remove pairs that have inappropriate content, such as articles that are neither related to COVID-19 nor the Vietnamese community or are too complicated to comprehend.

Second, at the pair-level, we check the spelling and correct mistakes. Undesirable artifacts, erroneous character and HTML tags will also be removed in this stage. In case the collected articles are too long or contains more than three paragraphs, extracted answers will be manually reduced in length by removing unnecessary information, such as source name, footnote, greetings, preliminary, over-specified explanations, and farewell. After this step, 4,500 Vietnamese question-answer pairs about COVID-19 are achieved from trusted FAQ sources. Some examples of the collected question-answer pairs are shown in Table 1.

3.3 Paraphrase Generation Process

Inspired by the concept of Kacupaj et al. (2021), we have investigated extending our dataset using the following paraphrasing methods on Vietnamese samples:

- Rearrange words, phrases, or sentences in the initial answer to create new responses without changing their meaning.
- Reduce or diversify the content of the initial answer.
- Paraphrase the initial answer using synonyms and similar structures.

These methods are manually applied on the collection to create up to three individual paraphrased responses consecutively for each question-answer pair. Newly created answers will be annotated in order to indicate the minimum number of answers per question.

After creating multiple paraphrased versions of the initial answer, we perform a cross-checking process to correct the spelling mistakes and grammatical errors as well as modify and rephrase digressive answers. By the end of this phase, the UIT-ViCoV19QA dataset is entirely constructed with 4,500 question-answer pairs containing at least one

Question: Xử lý triệu chứng ho khi chăm sóc F0 tại nhà thế nào? [**English:** How to handle cough symptoms when taking care of F0 patient at home?]

Answer: Dùng thuốc giảm ho theo đơn của bác sĩ. Có thể dùng thêm các vitamin theo đơn thuốc của bác sĩ. [**English:** Take cough suppressants as prescribed by your doctor. Additional vitamins can be taken according to the doctor's prescription.]

Question: Tôi đang điều trị viêm tiết niệu và viêm dạ dày hội chứng ruột kích thích thì có tiêm vaccine Covid-19 được hay không? [**English:** I am being treated for UTIs and gastritis with irritable bowel syndrome, can I get the Covid-19 vaccine?]

Answer: Với bệnh cấp tính mà anh/chị đang mắc phải cần được điều trị ổn định trước, sức khỏe tốt, bình thường thì có thể tiêm vaccine Covid-19. [**English:** If you have an acute illness that are stably treated, if you are in good health, you can receive the Covid-19 vaccine.]

Question: Hỗ trợ hô hấp cho trẻ em nhiễm COVID-19 ở thể nặng như thế nào? [**English:** How to provide respiratory support to children with COVID-19 in severe condition?]

Answer: Thở mask có túi Hoặc: NCPAP, HPNO, NIPPV [**English:** Apply breathing mask with bag or: NCPAP, HPNO, NIPPV]

Table 1: Examples of question-answer pairs from UIT-ViCoV19QA

answer and at most four unique paraphrased answers per question.

3.4 Statistics

The statistics of the training, development, and test sets are described in Table 2. The UIT-ViCoV19QA dataset consists of 4,500 question-answer pairs in total. In the table, the average length, as well as the vocabulary size [3] of questions and answers, are also presented.

Figure 2 illustrates the distribution of 4,500 questions of UIT-ViCoV19QA based on number of answers per question. The figure shows that the dataset contains 1800 questions that have at least two answers, 700 questions have at least three answers and half of them have a maximum of four paraphrased answers.

4 Methodologies

4.1 Baseline Models

This experimental section set up various deep learning models with Encoder-Decoder architecture to evaluate the UIT-ViCoV19QA dataset. These models have achieved significant results on many sequence-to-sequence learning tasks that involve long sequences, such as machine translation, text summarization, and question answering.

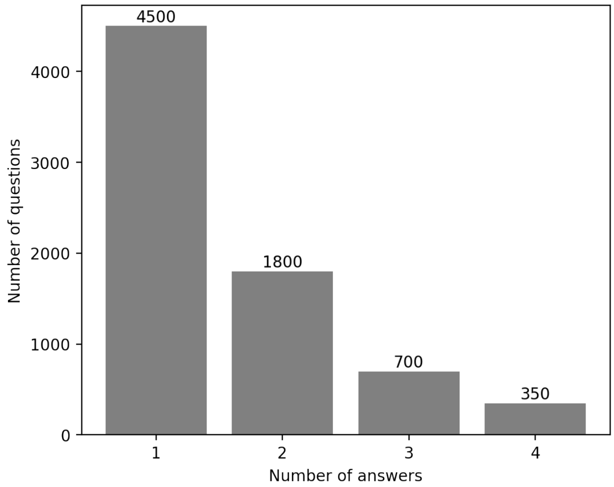


Figure 2: Distribution of total number of answers per question in UIT-ViCoV19QA

- **Attention-based Recurrent Neural Network**: Recurrent neural network (RNN) models used in the experiments are implemented with two attention mechanism - Bahdanau Attention (Bahdanau et al., 2016) and Luong Attention (Luong et al., 2015). To ensure that the results

[3] We use underthesea package: `https://github.com/undertheseanlp/underthesea` for word segmentation.

		Train	**Dev.**	**Test**	**All**
Answer 1	Number of question-answer pairs	3500	500	500	4500
	Average question length	31.44	33.66	32.32	31.79
	Average answer length	120.53	116.04	118.11	119.76
	Question vocabulary size	4396	1869	1770	4924
	Answer vocabulary size	8537	3689	3367	9411
Answer 2	Number of question-answer pairs	1390	209	201	1800
	Average question length	35.56	39.22	39.72	36.45
	Average answer length	40.54	39.25	42.73	40.64
	Question vocabulary size	2883	1269	1207	3305
	Answer vocabulary size	2632	1098	1129	2949
Answer 3	Number of question-answer pairs	542	79	79	700
	Average question length	34.77	36.70	39.28	35.49
	Average answer length	28.68	26.43	30.89	28.67
	Question vocabulary size	1836	717	693	2111
	Answer vocabulary size	1554	503	585	1753
Answer 4	Number of question-answer pairs	272	39	39	350
	Average question length	36.57	37.59	42.15	37.10
	Average answer length	29.75	29.03	35.72	30.25
	Question vocabulary size	1315	470	460	1519
	Answer vocabulary size	924	353	374	1075

Table 2: Overall statistics of the UIT-ViCoV19QA dataset.

are comparable, these models are set up with similar hyperparameters in the encoder and decoder as follows: an embedding layer with dimension 512, two hidden layers of 512 gated recurrent unit (GRU) cells, and a drop-out rate of 0.5. Bidirectional gated recurrent unit (BiGRU) is applied in the encoder of both models to help them understand the context better. For simplicity, RNN models using Bahdanau attention and Luong attention are annotated as RNN-1 and RNN-2, respectively.

- **Convolutional Network** (Gehring et al., 2017): Different from RNN, convolutional neural network uses many convolutional layers typically applied in image processing. Each layer uses filters to learn to extract different features from the text. In our experiments, the hyperparameters of models are set as follows: embedding layer with dimension 512, three convolutional layers with hidden size 512 use 1024 filters with kernel size 3 x 3, and drop-out probability 0.5.

- **Transformer** (Vaswani et al., 2017): As a dominant architecture in natural language processing (NLP), the model and its variants, such as BERT and pre-trained versions of BERT, have been commonly used to achieve state-of-the-art results for many tasks in the field. The model is set up with these settings: embedding layer with dimension 512, two layers with 8 self-attention heads, positional embedding layer with max length 500, position-wise feed-forward layer with dimension 2048 and drop-out rate 0.5.

4.2 Evaluation Metrics

Three standard metrics utilized for evaluating baseline models. These metrics are commonly used in machine translation and text summarization tasks to compare the generated text with human performance.

- **BLEU** (Papineni et al., 2002): BLEU is an n-gram based evaluation metric, widely used for Machine Translation (MT) evaluation to claim a high correlation with human judgments of quality. It aims to count the n-gram overlaps in the reference by taking the maximum count of each n-gram and clipping the count of the n-grams in the candidate text to the maximum count in the reference. In our experiments, BLEU score is calculated using unigram (BLEU-1) and 4-gram (BLEU-4) with uniform weight $w_n = 0.25$.

- **METEOR** (Lavie and Agarwal, 2007): The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. Moreover, it has several features not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. Unfortunately, those features are not yet supported for the Vietnamese language. We set up standard METEOR only using exact matching for evaluation with $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$.

- **ROUGE-L** (Lin, 2004): This metric measures the longest common subsequence (LCS) between our model output and reference. The idea here is that a longer shared sequence would indicate more similarity between the two sequences. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence-level word order. For each generated answer, we choose the best score achieved from comparing it with all existing reference answers.

4.3 Experimental Configuration

To assess the dataset's quality and illustrate the effect of having multiple paraphrased answer, we conduct and run individual experiment with four different dataset settings: one answer, two answers, three answers and finally, four paraphrased answers per question.

For Transformer, Adam optimizer is implemented with parameters as follows: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and $warmup\_step = 2000$. This setup varies the learning rate of the model during training progress by increasing it linearly for the first 2000 training steps and decreasing it after that proportionally to the inverse square root of the step number. For other models, standard Adam optimizer with a fixed learning rate of 0.001 is applied in the training.

The training progress is configured with a batch size of 8 in 30 epochs on an NVIDIA Tesla P100 GPU via the Kaggle platform[4]. After each epoch, the performance loss on the train and development sets is calculated using the Cross-Entropy Loss function. The maximum length of the model-generated output is limited to 500 tokens to reduce the generating time of repetitive loops caused by text degeneration.

5 Experimental Results

Table 3 presents the performance of our baseline models on different settings of the UIT-ViCoV19QA dataset. The final score of each metric in the table is achieved by calculating the average scores of all model-generated answers.

Evaluated using BLEU-1 and BLEU-4, the performance of models tends to improve when applying more paraphrased answers in the experiments, though this trend behaves differently among models. RNN-1 using two answers per question achieved the best BLEU-1 of 26.62% while on BLEU-4, Transformer using four answers outperforms other models with a score of 14.38%. Apart from RNN-1 achieves best score when applying two answers, other models have highest performance when training and evaluating using four answers.

In contrast with the BLEU score, the performance of models evaluated by METEOR and ROUGE-L varies significantly under different dataset settings. As shown in the table, RNN-2 and Convolutional perform well when using only one response, with scores of 33.95% and 32.29% respectively, while RNN-1 and Transformer need to apply more answers to achieve better scores.

Three examples of models generated responses are shown in Table 4 to illustrate the generation performance comparing with the original answer.

5.1 Error Analysis

From the experiment results, we determine that the Transformer trained on four answers gives the best

[4] https://www.kaggle.com/

Model	# answers	BLEU-1 (%)	BLEU-4 (%)	METEOR (%)	ROUGE-L (%)
RNN-1 (Bahdanau et al., 2016)	1	21.79	10.29	25.34	32.36
	2	**26.62**	**12.86**	25.15	33.66
	3	26.09	12.63	**25.98**	**33.68**
	4	24.56	12.27	23.91	32.57
RNN-2 (Luong et al., 2015)	1	21.04	10.94	**24.72**	**33.95**
	2	24.07	11.30	23.81	31.92
	3	23.8	10.95	23.65	31.65
	4	**24.38**	**12.30**	24.29	32.09
Convolutional (Gehring et al., 2017)	1	19.17	9.51	**22.76**	**32.29**
	2	21.63	10.81	20.66	30.71
	3	22.08	10.68	20.96	30.84
	4	**23.26**	**10.93**	21.84	31.15
Transformer (Vaswani et al., 2017)	1	21.84	10.82	23.37	31.67
	2	24.81	13.22	**24.60**	**32.25**
	3	24.90	13.64	21.72	32.20
	4	**25.19**	**14.38**	23.11	32.19

Table 3: Performance of baseline models on different dataset settings of UIT-ViCoV19QA

performance among others based on BLEU-4 according to the approach of Kacupaj et al. (2021). Twenty samples are randomly chosen from the generated answers of the model to perform error analysis. We calculate the average length and vocabulary size and count the number of POS tags in the references and generated answers. The statistics of the analysis process are shown in Table 5.

	Original answers	Generated answers
Avg. length	115.53	75.25
Vocabulary size	776	114
# POS tag	16	12

Table 5: The average length, vocabulary size and number of POS tag in 20 generated samples and their references

The vocabulary size of the output answers is almost seven times less than that of the original answers, while the average length and number of POS tags do not vary significantly. This implies that Transformer does not generalize well, and the tokens overlapping or text degeneration may have occurred in the outputs of the model.

By checking the generated results of Transformer, we notice some output answers contain overlapping phrases and do not match the question and reference answers. Further review the output of other models, various mistakes and errors have also been pointed out in the generated responses. In summary, these errors can be divided into three main groups.

- The first group includes meaningless and illogical answers. In some cases, the generated answers are just a set of tokens arranged in a chaotic order.
- The second group contains meaningful and coherent answers but do not satisfy the question's requirements, as illustrated by the second example in Table 4. This type of error can be caused by the diverse Vietnamese vocabulary and grammar or long passage, which makes the model difficult to understand the context of question.
- The third group consists of text degeneration. Typically, this is a common phenomenon in sequence-to-sequence learning tasks when a word or a phrase is infinitely repeated in the generated sequence indicating that the models may not generalize well. As illustrated by the third example in Table 4, the generated response does not satisfy the question

Question: tôi bị dị ứng với thuốc giảm đau giãn cơ . tôi uống thuốc này lúc bị sút lưng (phồng đĩa đệm) . vậy xin hỏi tôi có tiêm vaccine covid-19 được không ạ ? [**English:** I am allergic to muscle relaxants. I took this medicine when I had a low back pain (bulging disc). So can I get the covid-19 vaccine?]

Original Answer: chào anh / chị , nếu tình trạng dị ứng của anh / chị ở mức độ thông thường , không có phản vệ từ độ 2 trở lên , anh / chị vẫn có thể tiến hành tiêm chủng tại bệnh viện được . chúc anh / chị nhiều sức khỏe . trân trọng ! [**English:** hello sir / madam , if your allergy is at normal level , not having 2nd degree anaphylaxis or higher , you can still carry out vaccination at the hospital . Wish you a lot of health. Best regards !]

Generated Answer: chào anh , theo hướng dẫn của bộ y tế , anh thuộc nhóm đối tượng cần thận trọng khi tiêm chủng vaccine covid-19 . tuy nhiên , anh cần cung cấp đầy đủ thông tin về tình trạng dị ứng của mình cho bác sĩ khám sàng lọc để được tư vấn cụ thể hơn . cảm ơn câu hỏi của anh . trân trọng ! [**English:** hello sir, according to the guidance of the Ministry of Health, you are in the group of subjects that need to be cautious when vaccinating against the covid-19 vaccine. however, you need to provide enough information about your allergy to the screening doctor for more specific advice. Thanks for your question. best regards !]

Question: ai không nên tiêm vaccine covid-19 ? [**English:** who should not get the covid-19 vaccine?]

Original Answer: theo tổ chức y tế thế giới (who) , mỗi quốc gia cần chủng ngừa cho ít nhất 70 % dân số mới đạt được miễn dịch cộng đồng . giống với các loại dược phẩm khác , vaccine covid-19 có thể gây một số tác dụng phụ và không phù hợp ở từng đối tượng nhất định . tất cả những người có phản ứng dị ứng nghiêm trọng với bất kỳ thành phần hoạt chất hoặc tá dược nào liệt kê trong mục thành phần vắc xin đều không được tiêm vắc xin . bên cạnh đó , cần thận trọng tiêm chủng với các đối tượng , như : nếu bạn đã từng có phản ứng dị ứng nghiêm trọng (phản ứng phản vệ) sau khi tiêm vắc xin nào đó trước đây . nếu hệ miễn dịch của bạn bị suy yếu (suy giảm miễn dịch) hoặc bạn đang dùng thuốc làm suy yếu hệ miễn dịch (như corticosteroid liều cao , thuốc ức chế miễn dịch hoặc thuốc ung thư) . nếu bạn hiện đang bị nhiễm trùng nặng với thân nhiệt cao (trên 38 °c / 100.4 °f) . nếu bạn có vấn đề về xuất huyết / chảy máu hoặc bầm tím , hoặc nếu bạn đang dùng thuốc làm loãng máu (thuốc chống đông máu) . trong trường hợp nếu bạn không chắc chắn bất kỳ điều gì bên trên , hãy trao đổi với bác sĩ hoặc chuyên gia chăm sóc sức khỏe của bạn trước khi bạn được tiêm vắc xin . [**English:** according to the world health organization (who) , every country needs to immunize at least 70 % of the population to achieve herd immunity . Like other medicines, the covid-19 vaccine can cause some side effects and is not suitable for certain people. All people who have had a severe allergic reaction to any of the active ingredients or excipients listed in the vaccine ingredients section should not get the vaccine. In addition , care should be taken in vaccinating subjects , such as : if you have ever had a severe allergic reaction (anaphylaxis) after receiving a certain vaccine in the past . if your immune system is weakened (immunocompromised) or you are taking medicine that weakens your immune system (such as high-dose corticosteroids , immunosuppressants or cancer drugs) . if you currently have a severe infection with a high body temperature (above 38 °c / 100.4 °f) . if you have bleeding / bleeding or bruising problems , or if you are taking blood thinners (anticoagulants) . In case if you are not sure of any of the above, talk to your doctor or healthcare professional before you receive the vaccine.]

Generated Answer: người dân cần thực hiện nghiêm túc , chấp hành quy định của địa phương , thực hiện thông điệp 5k để bảo vệ bản thân và những người xung quanh .[**English:** people need to strictly comply with local regulations and implement the 5k message to protect themselves and those around them.]

Question: biến chủng Epsilon (B. 1.427 / B. 1.429) là gì ? [**English:** what is Epsilon (B. 1,427/B. 1,429) variant?]

Original Answer: biến chủng này có thể tự sắp xếp lại một phần của protein đột biến mà virus sử dụng để liên kết với các tế bào của cơ thể . vì thế , biến chủng epsilon có khả năng lây nhiễm cao hơn khoảng 20 % và dễ dàng làm suy yếu hệ miễn dịch của con người . [**English:** this variant can rearrange a portion of the mutated protein that the virus uses to bind to the body's cells. Therefore, epsilon variants are about 20% more contagious and easily weaken the human immune system.]

Generated Answer: theo quy định tại điều 11 nghị định số 176 / 2013 / nđ-cp ngày 14/11/2013 của chính phủ về xử phạt hành chính trong lĩnh vực y tế , hành vi không thực hiện biện pháp cách ly y tế tại điều 11 nghị định số 176 / 2013 / nđ-cp ngày 14/11/2013 của chính phủ về xử phạt hành chính trong lĩnh vực y tế , hành vi có hành vi có hành vi này bị phạt tiền từ 5.000.000 đồng đến 10.000.000 đồng đến 10.000.000 đồng đến 10.000.000 đồng ... [**English:** in accordance with Article 11 of the Government's Decree No. 176/2013/nd-cp dated November 14, 2013 on administrative sanctions in the health sector, failure to implement medical isolation measures in Article 11 of the Government's Decree No. 176/2013/nd-cp dated November 14, 2013 on administrative sanctions in the health sector, acts committed in this act are subject to a fine of from VND 5,000,000,000 to VND 10,000,000 to VND 10,000,000 to VND 10,000,000 ...]

Table 4: Examples of answers generated by the proposed models compared with the original answers

and the phrase "tới 10.000.000 đồng"(to VND 10,000,000) keeps repeated until the end of the answer.

6 Conclusion and Future Works

In this paper, we presented UIT-ViCoV19QA, the first community-based question answering dataset about COVID-19 for Vietnamese. Our dataset comprises 4,500 question-answer pairs with multiple paraphrased answers. The dataset's quality was evaluated through various baseline deep learning models and commonly used metrics such as BLEU, METEOR, and ROUGE-L.

We illustrated the effect of having multiple paraphrased answers for experiments with baseline models and provided benchmark results for further research. RNN with Bahdanau attention achieves the best BLEU-1 and METEOR scores of 26.62% and 25.98% when applying two and three answers respectively. Transformers using four answers outperforms others on BLEU-4 with score of 14.38%. On ROUGE-L, RNN with Luong Attention using one answer has the best performance of 33.95%. The advantage of having multiple paraphrased answers is greatly illustrated by BLEU scores, on which three out of four models achieve the best performance when applying all four paraphrased answers. On the contrary, our experiments showed that METEOR and ROUGE-L scores do not give a clear reflection of the improvement in models performance when increase number of answer used.

Through error analysis, we showed that the performance of these models is not quite good since the generated answers contain various errors. There are several reasons for this: the diversity of the Vietnamese language, lacking a specific evaluation metric for the Vietnamese language, the long-sequence content, and the size limitation of our dataset. The dataset offers a valuable contribution to the community, providing the foundation for many research lines in the single-turn QA domain and other areas.

In the future, UIT-ViCoV19QA can be expanded in size by collecting more relevant question-answer pairs and creating more paraphrased answers. The embedding layer of the proposed baselines can also be investigated to be replaced with pre-trained word embeddings for Vietnamese, such as PhoBERT (Nguyen and Nguyen, 2020), to improve model performance.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[Bahdanau et al.2016] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

[Choi et al.2018] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

[Del Tredici et al.2021] Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adrià de Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2974–2978.

[Gehring et al.2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

[Goyal et al.2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

[Hashemi et al.2020] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

[Kacupaj et al.2021] Endri Kacupaj, Barshana Banerjee, Kuldeep Singh, and Jens Lehmann. 2021. Paraqa: A question answering dataset with paraphrase responses for single-turn conversation.

[Lai et al.2017] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference* *on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Lavie and Agarwal2007] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June. Association for Computational Linguistics.

[Lin2004] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

[Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

[Luu et al.2021] Son T Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *International Conference on Computational Collective Intelligence*, pages 546–558. Springer.

[Nguyen and Nguyen2020] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

[Nguyen et al.2020] Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[Qiu and Huang2015] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth international joint conference on artificial intelligence*.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

[Reddy et al.2019] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

[Tran et al.2021] Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. Vivqa: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 546–554.

[Van Nguyen et al.2020a] Kiet Van Nguyen, Khiem Vinh Tran, Son T Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020a. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

[Van Nguyen et al.2020b] Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. New vietnamese corpus for machine reading comprehension of health news articles. *arXiv preprint arXiv:2006.11138*.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Zhang et al.2018] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Thirty-second AAAI conference on artificial intelligence*.

[Zhang et al.2021] Wei Zhang, Zeyuan Chen, Chao Dong, Wen Wang, Hongyuan Zha, and Jianyong Wang. 2021. Graph-based tri-attention network for answer ranking in cqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14463–14471.

Vowel sequences in Old Japanese: from a corpus-based approach

Chihkai Lin
National Taiwan University of Science and Technology
No.43, Keelung Rd., Sec.4, Da'an Dist.,
Taipei City 106335, Taiwan (R.O.C.)
linchihkai@gmail.com

Abstract

This paper investigates vowel sequences in Old Japanese from a corpus-based approach by using data from the Oxford-NINJAL Corpus of Old Japanese (ONCOJ). Three conditions are taken into consideration to analyze the corpus data, and they are (a) within a phonological word, (b) through morphological process, and (c) by grammatical element. The results have shown that vowel sequences in Old Japanese are not random combinations of vowels, but they are phonologically constrained. In the first and second conditions, the vowel sequences tend to show vowel$_{\text{[-front]}}$ + vowel$_{\text{[-low]}}$. In the third condition, the first vowel of the vowel sequence is -i-. In addition to frequency, this paper also discusses the types of words and compares types with frequencies.

1 Introduction

This paper investigates vowel sequences in Old Japanese from a corpus-based approach. One of the phonological characteristics of Old Japanese is that the basic syllable is open, CV, and V (Vovin, 2005; Fellesvig, 2010). While CV and V syllables are legitimate in Old Japanese, contact of two vowels generates a sequence like CV.V or V.V, which is uncommon in Old Japanese.

To avoid vowel sequences, a repair like vowel deletion, consonant insertion, or vowel coalescence becomes a possible strategy in Old Japanese. Much previous research has focused on vowel deletion (Oho, 1955, 1977; Yamaguchi, 1985; Unger, 1993; Kishida, 1998; Russell, 2003; Vovin, 2005; Frellesvig, 2010) and has tried to adopt different mechanisms, such as sonority scales and syllable numbers, to account for vowel deletion in Old Japanese. Although vowel deletion indeed takes place in Old Japanese, not all vowel sequences render vowel deletion, and vowel sequences still exist in Old Japanese. In general, the vowel sequences occur in three conditions, as shown in (1).

(1) a. within a phonological word;
b. through morphological process;
c. by grammatical element

In the first condition, consonant deletion in the intervocalic positions would lead to the contact of two vowels. For example, *makadi* 'paired oars' has a variant *makai* without the consonant *d*. Another example for the first condition is *mi-ato* HON-foot.print 'foot print'.[1] The formation of this word is a prefix of honorifics *mi-* and the word *ato* 'foot print'.

In the second condition, most examples are compounds. When the first word is followed by another word beginning with a vowel, a compound is formed and the two vowels contact, as in *opo-isi* 'big-stone'. In addition to phonological (1a) and morphological (1b) factors, grammatical element also renders vowel sequences in Old Japanese. For

[1] The abbreviations of the glosses in this paper are listed as follows: ATTR= attributive; CAUS = causative; CONC = concessive; COP = copular; DAT = dative; GEN = genitive; HON = honorifics; INF = infinitive; NEG = negation; PART = particle; PASS = passive; PERF = perfective; PFX = prefix; TOP = topic.

example, the infinitive marker in Old Japanese is vowel *-i-*, as in *k-i-ir-i* 'come-INF-enter-INF', and there is a vowel sequence *-i-i-* from the grammatical element *-i-* and the first vowel *i* in the root *-ir-* 'enter'.

Although the three conditions account for vowel sequences in Old Japanese, it remains unknown what the exact distributions of vowel sequences are in the three conditions and what the phonological constraints are, if there are, in Old Japanese. Therefore, before we explore phonological changes like vowel deletion, there is no doubt that we have to probe into the distributions of vowel sequences in Old Japanese. To achieve this goal, this paper adopts a corpus-based approach by collecting data from an online database, the Oxford-NINJAL Corpus of Old Japanese (ONCOJ). To facilitate the discussion, this paper is organized as follows. Section 2 introduces the Oxford-NINJAL Corpus of Old Japanese (ONCOJ) and discusses data selection criteria and analysis procedures. Section 3 reports the results, and section 4 discusses the phonological constraints in the three conditions. Besides, section 4 discusses the types of words in the corpus. Section 5 concludes this paper.

2 Corpus and data selection criteria

This section discusses the corpus used in this paper and data selection criteria. The data are collected from https://oncoj.ninjal.ac.jp/, with seven major philological sources, as listed in (2).

(2) a. *Kojiki kayō* (KK), 112 poems; 2,527 words. Compiled 712 CE
b. *Nihon shoki kayō* (NSK), 133 poems; 2444 words. Compiled 720 CE
c. *Fudoki kayō* (FK), 20 poems; 271 words. Compiled 730s CE
d. *Bussokuseki-ka* (BS), 21 poems; 337 words. Compiled after 753 CE
e. *Man'yōshū* (MYS), 4,685 poems; 83,706 words. Compiled after 759 CE
f. *Shoku nihongi kayō* (SNK), 8 poems; 134 words. Compiled 797 CE
g. *Jōgū shōtoku hōō teisetsu* (JSHT), 4 poems; 60 words. Date unknown

In ONCOJ, there are 4850 poems in total, and 97% of the poems are from MYS.

The steps of data collection are as follows. First, Old Japanese is transcribed in Chinese characters in two formats: phonographic and logographic. To unveil the phonetic values of the transcriptions, only the phonographic forms are analyzed in this paper. Second, vowels in ONCOJ are in the following notation: *a*, *e*, *o*, *i*, *u*, *wo*, *wi*, and *ye*. The eight vowels are conventionally divided into general vowels and paired vowels. The former includes vowels *a* and *u*; the latter includes pairs *i*/*wi*, *ye*/*e*, and *wo*/*o* (the vowels *i*, *ye*, and *wo* in the pair are known as *korui* vowels and the second vowels *wi*, *e*, and *o* in the pairs are known as *otsurui* vowels). The eight vowels lead to 64 possible combinations for vowel sequences. Third, the data are classified into three conditions depending on how the vowel sequence is recognized:

(a) in a phonological word, as in *kai* 'rudder' and *mi-ato* 'foot; foot print'
(b) compound, as in *opo-isi* 'big-stone';
(c) grammatical element, as in *k-i-ir-i* 'come-INF-enter-INF'.

Finally, since most of the data in Old Japanese are poems, this paper focuses on the situations where the vowel sequences occur in a phonological word. In other words, vowel sequences across a phonological word or a phrase, as in *wa-ga opo-kimi* 1st.pronoun-GEN big-lord 'my great lord' (MYS.18.4059), are not analyzed.

3 Results

In this paper, there are 627 examples of vowel sequences from ONCOJ, and Table 1 below shows the general distribution of the combinations of the eight vowels. In Table 1, the tendencies in the two vowels are different. In the first vowel, vowel *a* outnumbers other vowels. Vowels *i* and *o* are ranked second and third in Table 1. Other vowels, such as *u* and *wo*, take less than 10% of the total examples. The other three vowels *ye*, *e*, and *wi*, are rare in the corpus (< 5% in total). On the other hand, vowel *wo* is the majority in the second vowel, and vowel *ye* is in the second ranking. Fewer than 100 examples in the corpus, vowels *i*, *wi*, *a*, *o*, *u*, and *e*, less frequently appear in the second position of vowel sequence.

Table 1: The distribution of the combinations of the eight vowels

1st vowel 2nd vowel	a	i	u	ye	wo	e	o	wi	Total
a	7	38	0	0	0	2	1	1	49
i	10	28	12	1	0	6	7	6	70
u	11	10	2	1	2	0	0	0	26
ye	68	25	19	0	25	0	21	0	158
wo	115	23	14	4	14	0	56	0	226
e	0	1	0	0	0	3	0	0	4
o	3	28	1	0	0	0	0	0	32
wi	30	13	0	0	1	4	14	0	62
Total	244	166	48	6	42	15	99	7	627

In the 64 possible combinations of the vowel sequences, the cell that exceeds 100 examples is that of the combination of vowels *a* and *wo*, with 115 examples (18%, 115/627), as in *sa-wo-wo-ni pa* PFX-small-peak-DAT TOP 'to the small peak' (KK 89), and *awo-ni yo-si* red-clay good-COP 'good red clay' (MYS.5.806).

There are two cells with more than 50 examples. The first cell is that of the combination of vowels *a* and *ye*, as in *kasuka-ye-no* KASUKA-bay-GEN 'of Kasuka Bay' (KK 95); the second cell is that of the combination of vowels *o* and *wo*, as in *opo-wo-ni pa* big-peak-DAT TOP 'to the big peak' (KK 89). The following subsections discuss the distributions of vowel sequences in three categories: in a phonological word (3.1), compounds (3.2), and grammatical element (3.3).

3.1 Vowel sequences in a phonological word

The first category of vowel sequences is that in a phonological word. Words with affix, prefix in particular, are also included in this category. In total, there are 305 examples in this category, as shown in Table 2 below.

In Table 2, if we focus on the first vowel, more than half of the examples are in vowel *a* (157 examples). On the other hand, vowel *wo* is the majority in the second vowel (140 examples). In Table 2, the vowel *a* tends to cooccur with vowel *wo* or vowel *ye*, as in *awo-ni yo-si* red-clay good-COP 'good red clay' (MYS.5.806) and *aye-n-u gani* fall.off-PERF-ATTR PART 'fallen off' (MYS.8.1507). Other frequent combinations like vowels *a* and *wi* are also found in the corpus, as in *kurenawi n-i* safflower COP-INF 'the safflower' (MYS.18.4111). Other first vowels with sufficient examples are vowel *o* (52 examples) and vowel *wo* (40 examples), as in the conjunctional particle *monowo*. There are some sporadic examples such

Table 2: The distribution of the combinations of the eight vowels (in a phonological word)

1st vowel 2nd vowel	a	i	u	ye	wo	e	o	wi	Total
a	2	9	0	0	0	0	0	0	11
i	8	0	6	0	0	0	0	4	18
u	7	2	0	0	1	0	0	0	10
ye	51	3	11	0	26	0	8	0	99
wo	68	8	7	0	13	0	44	0	140
e	0	0	0	0	0	1	0	0	1
o	2	4	0	0	0	0	0	0	6
wi	19	1	0	0	0	0	0	0	20
Total	157	27	24	0	40	1	52	4	305

as *kwoye-te k-ite* pass.over-GEN come-GEN 'pass over and come' (MYS.15.3762).

3.2 Vowel sequences in compound

In the second category of vowel sequences, there are 131 examples in the corpus. Table 3 shows the distribution of the combinations of the eight vowels in compounds.

Some other sporadic examples are also found in the corpus, such as vowels *o* and *wi* (13 examples) in *kumo-wi nasu* cloud-be.at COP 'clouds' (MYS.17.4003), and vowels *u* and *ye* (8 examples) in *sidu-ye pa* lower-branch TOP 'lower branch' (KK 43).

Table 3: The distribution of the combinations of the eight vowels (in compounds)

1st vowel 2nd vowel	a	i	u	ye	wo	e	o	wi	Total
a	5	0	0	0	0	2	1	0	8
i	2	0	3	1	0	5	3	0	14
u	4	0	1	1	1	0	0	0	7
ye	4	0	8	0	0	0	0	0	12
wo	47	5	7	4	1	0	4	0	68
e	0	0	0	0	0	2	0	0	2
o	1	0	1	0	0	0	0	0	2
wi	3	1	0	0	1	0	13	0	18
Total	66	6	20	6	3	9	21	0	131

If we focus on the first vowel in Table 3, approximately half of the examples are in vowel *a* (66 examples). Likewise, the majority in the second vowel is vowel *wo* (68 examples).

The commonest vowel sequence in Table 3 is that of vowels *a* and *wo* (47 examples), as in *masura-wo no* strong-male COP 'strong man' (MYS.5.804).

3.3 Vowel sequence in grammatical element

In the third category of vowel sequences, there are 191 examples in the corpus, as shown in Table 4 below.

Different from the distributions in Tables 2 and 3, the one in Table 4 shows that the majority of the first vowel goes to vowel *i* (70%, 133/191), as in *wor-i-ak-as-i mo* sit-INF-bright-CAUS-INF PART

Table 4: The distribution of the combinations of the eight vowels (in grammatical element)

1st vowel 2nd vowel	a	i	u	ye	wo	e	o	wi	Total
a	0	29	0	0	0	0	0	1	30
i	0	28	3	0	0	1	4	0	36
u	0	8	1	0	0	0	0	0	9
ye	13	22	0	0	0	0	12	0	47
wo	0	9	0	0	0	0	8	0	17
e	0	1	0	0	0	0	0	0	1
o	0	24	0	0	0	0	0	0	24
wi	8	12	0	0	0	4	1	2	27
Total	21	133	4	0	0	5	25	3	191

‘sit up all night’ (MYS.18.4068) or in *yeda kir-i-oros-i* branch cut-INF-let.down-INF ‘cut down the branches’ (MYS.15.3603). Other vowel sequences are mostly attested when the second vowel is *ye*, as in *i-no ne-raye-n-u ni* sleep-GEN sleep-PASS-NEG-ATTR COP ‘(I) do not sleep’ (MYS.15. 3665), and *kik-o-ye sikado mo* hear-INF-PASS CONC PART ‘Although I am heard,’ (KK 45).

Thus far, we have discussed the distributions of the vowel sequences in the corpus, and the results have shown that the vowel sequences in the first and second conditions in (1) differ from those in the third condition in (1).

4 Discussion

This section discusses phonological constraints in the vowel sequences in Old Japanese (4.1) and the words with the vowel sequences (4.2).

4.1 Phonological constraints

The three categories of vowel sequences in the corpus show two phonological constraints in Old Japanese. In section 3, we have discussed the three categories of vowel sequences in Old Japanese. Of the three categories, the third category is different from the other categories, as the first vowel in high frequency is an infinitive marker, a high front vowel *i* (vowel *o* as a variant of vowel *i*).

If we temporarily remove the distribution of grammatical element from Table 2, the combined distribution of a phonological word and compound, as shown in Table 5 below, reflects the first phonological constraint of the vowel sequences in Old Japanese.

The five vowel sequences that exceed 5 % of the corpus examples are *a* + *wo* (115 examples, 26%), *a* + *ye* (55 examples, 13%), *o* + *wo* (48 examples, 11%), *wo* + *ye* (26 examples, 6%), *a* + *wi* (22 examples, 5%).

The vowel sequences in Table 5 reflect the ranking of the unmarkedness of the vowels in the two positions: *a* > *o* > *u* > *wo* > *i* > *e* > *wi* for the first vowel and *wo* > *ye* > *wi* > *i* > *a* > *u* > *o* > *e* for the second vowel. It is apparent that vowel sequences in Old Japanese are not just random combinations of any two vowels. The vowel sequences are preferred to be vowel$_{\text{[-front]}}$ plus vowel$_{\text{[-low]}}$.

The second phonological constraint based on the corpus data is that the general vowels, *a* and *u*, and the paired vowels *i/wi*, *ye/e*, and *wo/o* (*korui* vowels including *i*, *ye* and *wo*, and *otsurui* vowels including *wi*, *e*, and *o*) are positionally different in the vowel sequences. Table 6 below shows the nine combinations of general and paired vowels.

The combination of general and *korui* vowels outnumbers other combinations in the corpus. The data in Table 6 suggest that the first vowel is preferred to be the general vowels, and the second vowel is preferred to be the *korui* vowels.

Table 5: The distribution of the combinations of the eight vowels (without grammatical element)

1st vowel 2nd vowel	a	i	u	ye	wo	e	o	wi	Total
a	7	9	0	0	0	2	1	0	19
i	10	0	9	1	0	5	3	4	32
u	11	2	1	1	2	0	0	0	17
ye	55	3	19	0	26	0	8	0	111
wo	115	13	14	4	14	0	48	0	208
e	0	0	0	0	0	3	0	0	3
o	3	4	1	0	0	0	0	0	8
wi	22	2	0	0	1	0	13	0	38
Total	223	33	44	6	43	10	73	4	436

Table 6: Combinations of general and paired vowels

1st vowel / 2nd vowel	G	K	O	Total
G	19	14	3	36
K	222	61	68	351
O	26	7	16	49
Total	267	82	87	436

*G = general vowels; K = *korui* vowels; o = *otsurui* vowels

To sum up, the two phonological constraints in vowel sequences in Old Japanese are vowel$_{[\text{-front}]}$ plus vowel$_{[\text{-low}]}$ in terms of vowel types and general vowel plus *korui* vowel in the terms of vocalic categories.

4.2 The words with the vowel sequences

In this section, we discuss the vowel sequences in the word differences in the corpus. In Table 5, we have shown the frequencies of vowel sequences in the corpus, while some words are frequently used in the corpus. For example, the word *monowo* 'focus particle' is attested 38 times in the corpus, and the word *awo* 'green or blue' is attested 32 times. The fact that two words are highly frequently used in the corpus makes vowel *wo* prominent in the second vowel. To fully understand the vowel sequences in Old Japanese, it is necessary to calculate the vowel sequences in terms of types of words, not the frequency of words.

Table 7 shows distributions in a phonological word without affixation. Examples of compounds are also not included in the discussion. In Table 7, there are 50 words, and half of the words are those with vowel *a* as the first vowel, as in *taye* 'break' and *sakaye* 'prosper' for *a* + *ye* (9 words), *awo* 'green or blue' and *mawos-* 'speak [hum]' for *a* + *wo* (6 words), and *sawi* 'rushing' for *a* + *wi* (4 words).

If we consider the number of words together with the tokens in the vowel sequences in Table 7, the top three vowel sequences are *a* + *wo*, *a* + *ye*, and *o* + *wo*.

5 Conclusion

This paper has discussed the vowel sequences in Old Japanese from a corpus-based approach. This paper has discussed the vowel sequences in terms of three conditions: in a phonological word, in compound, and in grammatical element. The results have shown that the distribution of grammatical element differs from those of the other two categories in the vowel types since most examples in the category of grammatical element are reflected by the infinitive marker, vowel *i*.

Table 7: The number of words in vowel sequences

Vowel 1	Vowel 2	Words	Tokens
a	ye	9	50
a	wo	6	55
a	wi	4	19
i	wo	4	7
u	wo	4	7
o	wo	4	44
a	u	3	7
u	ye	3	11
o	ye	3	8
wo	ye	2	26
a	a	1	2
a	i	1	8
i	ye	1	1
i	wi	1	1
u	i	1	6
wi	i	1	4
wo	wo	1	12
e	e	1	1
Total		50	269

On the basis of the corpus data, this paper has also discussed two phonological constraints and the types of vowel sequences in words. On the one hand, the first vowel tends to be [-front] and the second vowel to be [-low]; on the other hand, the first vowel tends to be general vowels and the second to be *korui* vowels. In addition to the frequencies of the corpus data, this paper has explored the words in different vowel sequences. The comparison of words and tokens has indicated that the most frequently used vowel sequences are *a* + *ye*, *a* + *wo*, and *o* + *wo*.

Much effort has been made to discuss the vowel sequences within the domain of a phonological word in Old Japanese. It is also possible that vowel sequences are attested beyond the phonological word. For example, in *wa-ga opo-kimi* 1st.pronoun-GEN big-lord 'my great lord' (MYS.18.4059), there is a vowel sequence between the genitive case marker *ga* and the noun *opo-kimi*. In future research, one topic that needs in-depth investigation could be the vowel sequences across phonological words. After that, a complete picture of the vowel sequences in Old Japanese can be obtained.

Acknowledgments

A research grant from Taiwan's National Science and Technology Council 110-2410-H-011-033- is gratefully acknowledged. The usual disclaimer applies.

References

Alexander Vovin. 2005. A Descriptive and Comparative Grammar of Western Old Japanese. Global Oriental Folkestone, Kent, UK.

Bjarke Frellesvig. 2010. A History of the Japanese Language. Cambridge University Press, Cambridge, UK.

Kelly Russell L, 2003. 'Contraction and Monophthongization in Old Japanese' in Toshiki O, Vovin A (ed.), Nihongo keitōron no genzai/Perspectives on the Origins of the Japanese Language, Nichibunken, Tokyo, Japan.

Marshall Unger, J. 1993 [1977]. Studies in early Japanese morphophonemics (2nd revised edition). Indiana University Linguistics Club Publications Bloomington, USA.

Susumu Ono. 1953. Manyoujidai no Onyin [Phonology in Manyou period], in Manyoushu Daisei 6: gengohen [Collection of Manyoushu 6: linguistics]. Heiponsha, Tokyo, Japan.

Susumu Ono. 1977. Onyin no Hensen 1 [Phonological changes 1], in Iwanami Kouza: Nihongo 5: Onyin [Seminar of Iwanami: Japanese phonology]. Iwanami, Tokyo, Japan.

Takeo Kishida. 1998. Kokugo Onyin Henkaron [A Study of Japanese Phonology]. Musashinoshoin, Tokyo, Japan.

Yoshinori Yamaguchi. 1985. Kodai Nihongo Bunbo no Seiritsu no Kenkyu [A Study of Old Japanese Grammar]. Yuseitou, Tokyo, Japan.

Effectiveness Analysis of Word Sense Disambiguation Using Examples of Word Senses from WordNet

Hiroshi Sekiya
Graduate School of Sci. and Eng., Ibaraki Univ. / 4-12-1 Nakanarisawacho, Hitachi City, Ibaraki Prefecture, 316-8511, Japan
22nm727x@vc.ibaraki.ac.jp

Minoru Sasaki
Graduate School of Sci. and Eng., Ibaraki Univ. / 4-12-1 Nakanarisawacho, Hitachi City, Ibaraki Prefecture, 316-8511, Japan
minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract

In Word Sense Disambiguation (WSD), several studies have proposed systems that incorporate dictionary glosses. The use of glosses expressed in short sentences can improve the accuracy of the WSD system. However, since many glosses are short sentences, additional lexical information could improve accuracy. In this study, we propose a method to incorporate examples of word senses described in WordNet 3.0 as new lexical information into BEM, a WSD system, and analyze the effectiveness of the examples of word senses. Specifically, examples of word senses are input into BEM, and the [CLS] vector or target word vector is taken from the output word embedding representation sequence and used as the sense embedding representation. In the experiment, out of six evaluation sets, including the development set, the F1 score decreased in five evaluation sets and improved in one evaluation set. Since the F1 score improved in one evaluation set, we expect that the use of examples of word senses would be effective.

1 Introduction

Word sense disambiguation (WSD) is one of the major tasks in natural language processing (NLP). WSD is the process of identifying the most appropriate sense for a polysemous word in context. This technique is crucial in many applications in other areas of NLP, such as machine translation (Nguyen et al., 2018), information extraction (Chai and Biermann, 1999), text summarization (Rahman and Borah, 2020), and so on.

Previous research has shown that incorporating lexical information, such as glosses, into a WSD system improves accuracy (Luo et al., 2018a, b; Blevins and Zettlemoyer, 2020). Glosses have been found to be effective for both most frequent sense (MFS) and less frequent sense (LFS). However, many of the glosses are written in short sentences, and it is not clear whether the glosses effectively capture the information in the sense. We expect that the use of information on many senses, rather than just glosses, will capture the characteristics of senses.

To solve this problem, we propose a WSD method to use examples of word senses from WordNet 3.0 (Miller, 1995) as additional lexical information in BEM (Blevins and Zettlemoyer, 2020). In the proposed system, the target word vector or [CLS] vector of examples of word senses is used as the sense embedding representations. We expect that the use of examples of word senses as well as glosses can effectively capture the characteristics of word senses. We compare the performance of this proposed method with that of BEM to test the effectiveness of examples of word senses in WSD.

2 Related Work

There are two types of word sense disambiguation (WSD) tasks: the Lexical Sample Task, in which the target words for WSD are predefined, and the All-words WSD, in which all polysemous words in a sentence are target words. This study is categorized as an All-words WSD.

The BEM by Blevins et al. consists of a context encoder that represents the target word and its surrounding context and a gloss encoder that represents the sense glosses, representing the target words and senses in the same embedding space. These two encoders are initialized with a pre-trained model, BERT (Devlin et al., 2019), and are jointly fine-tuning. This method outperforms the results for All-words WSD in English presented in the previous study by Raganato et al. (2017b). In this study, the model was trained in BEM by creating a new sense embedding representation of examples of word senses.

It has been shown that lexical information such as glosses is a valuable resource for improving the accuracy of WSD. Lesk (1986) used overlap between sense glosses and the context of the target word to estimate the sense of the target word. This method was later extended to incorporate WordNet graph structure (Banerjee and Pedersen, 2003). It has also been extended to incorporate word embeddings (Basile et al., 2014). In a recent study, Luo et al. (2018a, b) input sense glosses into a neural WSD system and significantly improved accuracy.

3 BEM

In this section, we introduce the BEM proposed by Blevins et al. The model structure of the BEM is shown in Figure 1. BEM is a supervised WSD system designed to efficiently utilize sense glosses that define a less frequent sense. BEM is composed of two independent encoders: a context encoder that represents the target word and surrounding context, and a gloss encoder that embeds the sense glosses. Each encoder is a deep transformer network initialized with BERT to take advantage of the word sense information obtained from prior training (Coenen et al., 2019; Hadiwinoto et al., 2019). Thus, the input to each encoder is padded with BERT-specific start [CLS] and end [SEP] symbols.

BEM is designed to encode contextualized target words and sense glosses independently (Bromley et al., 1994; Humeau et al., 2019), and each of these models is initialized with a BERT-base.

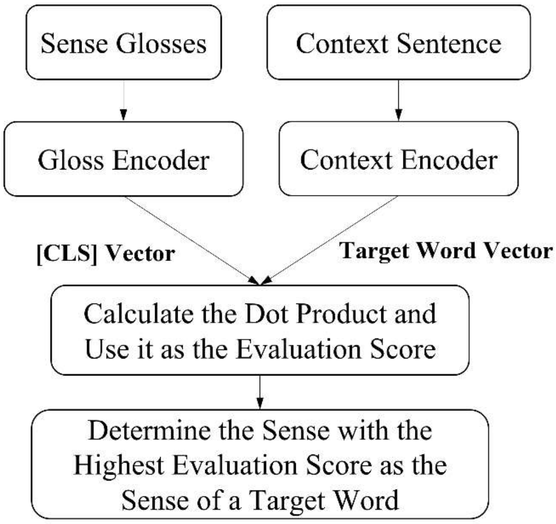

Sense Glosses
Context Sentence
Gloss Encoder
Context Encoder
[CLS] Vector
Target Word Vector
Calculate the Dot Product and Use it as the Evaluation Score
Determine the Sense with the Highest Evaluation Score as the Sense of a Target Word

Figure 1. Model Structure of BEM

The context encoder T_c takes as input a context sentence c containing the target words w for WSD. c is represented by $c = c_0, c_1, c_2, \cdots, w_i, \cdots, c_n$, where w_i is the i^{th} target word of the context sentence. The context encoder outputs a context-aware word embedding representation sequence r. The target word vector in the word embedding sequence r is denoted by r_{w_i}, where r_{w_i} is the i^{th} representation output by T_c. r_{w_i} is given by

$$r_{w_i} = T_c(c)[i]$$

For words tokenized into multiple subword by the BERT tokenizer, the word is represented by the average representation of the subword parts. For example, if the j^{th} through k^{th} tokens correspond to the subword of the i^{th} word, r_{w_i} is given by

$$r_{w_i} = \frac{1}{k-j} \sum_{l=j}^{k} (T_c(c)[l])$$

The gloss encoder T_g takes as input the gloss $g_s = g_0, g_1, \cdots, g_m$ that define the sense s. The [CLS] vector, which is the first representation in the word embedding representation sequence output by gloss encoder, is the global representation of s. The global representation of s is denoted by r_s, and r_s is given by

$$r_s = T_g(g_s)[0]$$

As shown in the following equation, each candidate sense $s \in S_w$ of the target word w is given a score by taking the dot product of r_w and every r_s.

$$\phi(w, s_i) = r_w \cdot r_{s_i}$$

where i is $i = 0, \cdots, |S_w|$. When evaluating, the meaning $\hat{s}$ of the target word w is predicted to be

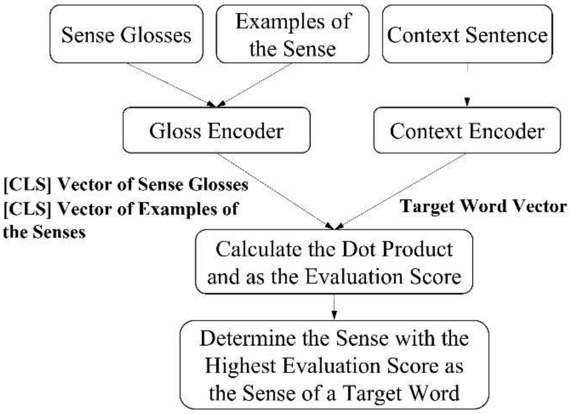

Sense Glosses
Examples of the Sense
Context Sentence
Gloss Encoder
Context Encoder
[CLS] Vector of Sense Glosses
[CLS] Vector of Examples of the Senses
Target Word Vector
Calculate the Dot Product and as the Evaluation Score
Determine the Sense with the Highest Evaluation Score as the Sense of a Target Word

Figure 2. Structure of the Model Using the [CLS] Vector of Examples of Word Senses

the $s_i \in S_w$ with the highest score for the dot product of r_{s_i} and r_w.

For the score of each candidate sense of the target word w, the BEM is trained using a loss function, cross-entropy loss. Given a word-sense pair (w, s_i), the loss function of the system is given by

$$\mathcal{L}(w, s_i) = -\phi(w, s_i) + \log \sum_{j=0}^{|S_w|} \exp\left(\phi(w, s_i)\right)$$

4 Method

In this study, we input examples of word senses into the BEM to estimate the sense of the target word. Specifically, we use the [CLS] vector and the target word vector obtained from the examples of word senses as the sense embedding representations. Three main types of examples of word senses representations are used to train the BEM, and three types of models are created using the examples of word senses representations. The method of representation of examples of word senses is explained in detail in Sections 4.1, 4.2, and 4.3. The examples of word senses are obtained from WordNet3.0.

4.1 Word Sense Disambiguation Using the [CLS] Vector of Examples of Word Senses

In this section, we present an approach using [CLS] vector of examples of word senses. The structure of the model using the [CLS] vector of examples of word senses is shown in Figure 2. If there are multiple examples of the sense for a sense, one of them is obtained. The obtained examples of the sense are input to the gloss encoder, and the [CLS] vector is extracted from the word embedding representation sequence output from the gloss encoder. We use the extracted [CLS] vector as the sense embedding representation. We take the dot product of the target word vector of the context sentence output by the context encoder and the [CLS] vector of the sense glosses output by the gloss encoder. In addition, we take the dot product of the target word vector of the context sentence output by the context encoder and the [CLS] vector of the examples of the sense output by the gloss encoder. The results of the dot product calculation are used as the score of each candidate sense. Then, the sense with the highest score among each candidate sense is estimated as the sense of target word.

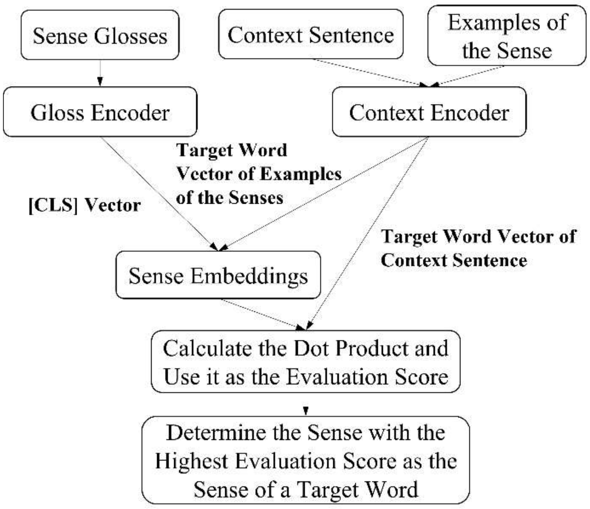

Sense Glosses
Context Sentence
Examples of the Sense
Gloss Encoder
Context Encoder
Target Word Vector of Examples of the Senses
[CLS] Vector
Target Word Vector of Context Sentence
Sense Embeddings
Calculate the Dot Product and Use it as the Evaluation Score
Determine the Sense with the Highest Evaluation Score as the Sense of a Target Word

Figure 3. Structure of the Model Using the Target Word Vector of Examples of Word Senses

4.2 Word Sense Disambiguation Using the Target Word Vector of Examples of Word Senses

In this section, we present an approach using the target word vector of examples of word senses. The structure of the model using the target word vector of examples of word senses is shown in Figure 3. If there are multiple examples of the sense containing the target word for a sense, one of them is obtained. For example, if the target word is "review," the examples of the sense with words of the same type as "review" in the sentence are obtained. The

Dataset	Part-of-speech of the Target Word	Number of Senses	Number of Annotated Example
SE07	Nouns, Verbs	375	455
SE2	Nouns, Verbs, Adj., Adv.	1335	2282
SE3	Nouns, Verbs, Adj., Adv.	1167	1850
SE13	Nouns	827	1644
SE15	Nouns, Verbs, Adj., Adv.	659	1022

Table 1. Characteristics of the Senseval/SemEval Dataset

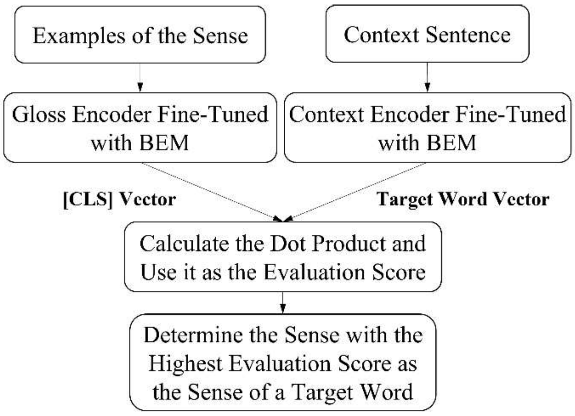

Examples of the Sense
Context Sentence
Gloss Encoder Fine-Tuned with BEM
Context Encoder Fine-Tuned with BEM
[CLS] Vector
Target Word Vector
Calculate the Dot Product and Use it as the Evaluation Score
Determine the Sense with the Highest Evaluation Score as the Sense of a Target Word

Figure 4. Structure of the Model with Vector of Sense Glosses Updated with Examples of Word Senses

obtained examples of the sense are input to the context encoder, and the target word vector is extracted from the word embedding representation sequence output from the context encoder. We use the extracted target word vector as the sense embedding representation. We take the dot product of the target word vector of the context sentence output by the context encoder and the [CLS] vector of the sense glosses output by the gloss encoder. In addition, we take the dot product of the target word vector of the context sentence output by the context encoder and the target word vector of the examples of the sense output by the context encoder. The results of the dot product calculation are used as the score of each candidate sense. Then, the sense with the highest score among each candidate sense is estimated as the sense of target word.

4.3 Word Sense Disambiguation with the vector of sense glosses updated with examples of word senses

In this section, we present an approach that updates the vector of sense glosses with examples of word senses. The structure of the model with the vector of the sense glosses updated with the examples of word senses is shown in Figure 4. If there are multiple examples of the sense for a sense, one of them is obtained. If no examples of the sense exist for a sense, the sense gloss is used as the examples of the sense as a substitute. The obtained examples of the sense are input to the gloss encoder, and the [CLS] vector is extracted from the word embedding representation sequence output from the gloss encoder. We use the extracted [CLS] vector as the sense embedding representation. We then update the word sense vector of the sense glosses by fine tuning each encoder with the sense glosses, followed by fine tuning with the examples of the sense.

5 Experiments

5.1 Datasets

We use the WSD framework created by Raganato et al. (2017b) to evaluate model performance. As training data, we use the SemCor corpus (Miller et al., 1993), a large dataset of manually annotated word senses from WordNet. SemCor contains 226,036 annotated examples covering 33,362 senses. We use the SemEval-2007 (SE07) dataset (Pradhan et al., 2007) as the development set. As evaluation sets, we use the Senseval-2 (SE2; Palmer et al. (2001)), Senseval-3 (SE3; Snyder and Palmer (2004)), SemEval-2013 (SE13; Navigli et al. (2013)), SemEval- 2015 (SE15; Moro and Navigli (2015)), and ALL datasets. The ALL dataset is a dataset that concatenates all development and evaluation sets. In addition, all sense glosses and examples of word senses used in this system are taken from WordNet 3.0.

The Senseval/SemEval dataset is a dataset focused on the WSD task. The characteristics of each dataset are shown in Table 1.

	SE07	SE2	SE3	SE13	SE15	ALL
BEM	**74.5**	**79.4**	77.4	**79.7**	**81.7**	**79.0**
BEM1	73.8	78.6	76.7	77.4	80.6	77.8
BEM2	73.6	79.3	**78.1**	78.2	80.3	78.5
BEM3	73.0	77.6	76.2	76.5	80.1	77.0

Table 3. F1 Score for Each Model

Model	Explanation
BEM1	Use [CLS] vector of examples of word senses (Section 4.1)
BEM2	Use target word vector of examples of word senses (Section 4.2)
BEM3	Vector of sense gloss updated with examples of word senses (Section 4.3)

Table 2. The Model Proposed in this Study

5.2 Experimental Setup

The models presented in Sections 4.1, 4.2, and 4.3 are shown in Table 2. We compare the F1 score of these models with the F1 score of the BEM to analyze the effectiveness of the examples of word senses. BEM1 and BEM3 are trained in the Google Colaboratory Pro+ and BEM2 are trained in the Google Colaboratory Pro environment.

Each model proposed in this study is trained in the Google Colaboratory environment, which has a limited runtime, and therefore epochs that can be done in a single run is limited. Also, each model saves only the model with the highest F1 score and resumes training. Therefore, if current run does not obtain a higher F1 score than the previous run, the model training is terminated at that point. As a result, epochs vary from model to model. BEM1 is trained 14 epochs, BEM2 is trained 6 epochs, and BEM3 is trained 17 epochs. We use context batch size 4; BEM1 and BEM2 use gloss batch size 128 and BEM3 uses gloss batch size 256.

5.3 Evaluation Method

We use the development and evaluation sets presented in Section 5.1 to evaluate the performance of our models by determining the F1 score of each model. We used the model with the highest F1 score in the development set to obtain the F1 score in the evaluation set.

	Zero-shot Words
BEM	91.2
BEM1	92.2
BEM2	**92.9**
BEM3	90.6

Table 4. F1 Score of Zero-shot Words

5.4 Results

The F1 score for BEM, BEM1, BEM2, and BEM3 on the development set and the five evaluation sets are shown in Table 3. BEM1, BEM2, and BEM3 with the use of examples of word senses resulted in lower F1 score than BEM in all development and evaluation sets except the Senseval-3 evaluation set. For the Senseval-3 evaluation set, the highest score is 78.1% for BEM2, which is a 0.7% improvement over BEM.

The F1 score for zero-shot words in the ALL evaluation set are shown in Table 4. zero-shot words are words that did not appear in the training data. BEM3 results in an F1 score below that of BEM, while BEM1 and BEM2 outperform the F1 score of BEM by up to 1.7%.

6 Discussion

The model using the examples of word senses results in lower F1 score than the original BEM, except for the senseval-3 evaluation set. This indicates that the examples of word senses are noise and have a negative impact on the system. However, since the senseval-3 evaluation set improves the F1 score, we expect that the use of examples of word senses may be effective if we devise a way to use them.

Among the models using examples of word senses, BEM2 shows the best results. therefore, we can say that the most effective way to incorporate examples of word senses into a model is to represent the examples of word senses as the target word vector. The reason for the best results with the target word vector is thought to be that unlike the sense glosses, the examples of word senses the usage of

	BEM-correct	BEM-wrong
BEM1-correct	-	249
BEM1-wrong	333	-
BEM2-correct	-	215
BEM2-wrong	249	-
BEM3-correct	-	251
BEM3-wrong	391	-

Table 5. Comparison of the number of correct and incorrect senses between BEM and BEM1~3. For example, 249 is the number of senses that are wrong in BEM but correctly estimated in BEM1.

the sense, not the meaning of the sense. This may have resulted in the [CLS] vector having a low F1 score because the [CLS] vector could not effectively express the features of the sense meaning due to the presence of many words in the examples of word senses that have a low similarity to the sense meaning. BEM2 also has a higher F1 score than the original BEM on the senseval-3 evaluation set. Therefore, we expect that the F1 score will be improved by devising the use of examples of word senses. For example, if the target word is "review," we are considering using "reviewed," the past tense of review, or "reviews," the plural of review, as the target word.

We find that BEM1 and BEM2 have a higher F1 score for zero-shot words than the original BEM. Therefore, we think that using the examples of word senses as sense embedding representations can effectively represent the features of words that do not appear in the training data.

To analyze the experimental results of this study in detail, we examined the number of senses that were wrong in the BEM but correct in the model using examples of word senses, and the number of senses that were correct in the BEM but wrong in the model using examples of word senses, based on the estimation results of the ALL evaluation set. The results of the survey are shown in Table 5. The survey results show that the proposed system can correctly estimate senses that are incorrectly estimated by BEM. However, more than that number, the proposed system incorrectly estimates senses that are correctly estimated by BEM. In particular, BEM1 and BEM3 incorrectly estimate 84 and 140 more than the original BEM, respectively. In contrast, BEM2 is estimated with 34 more errors than the original BEM. This indicates that although BEM2 has fewer correctly estimated senses than BEM1 and BEM3, it has less negative impact on the system than BEM1 and BEM3. These results indicate that although the examples of word senses have a negative impact on the system, there are many cases where a sense that is wrong in the original BEM is correctly estimated by the BEM using the examples of word senses. Therefore, we anticipate that the system's accuracy could be improved by devising ways to use examples of word senses.

	Nouns	Verbs	Adj.	Adv.
BEM1	4.49%	**5.39%**	4.40%	2.60%
BEM2	3.26%	**4.30%**	3.14%	2.31%
BEM3	5.02%	**7.20%**	4.50%	3.76%

Table 6. Percentage of Parts of Speech of Newly Mistaken Word Senses When Examples of Word Senses are Used

	Nouns	Verbs	Adj.	Adv.
SE07	2.52%	**6.42%**	-	-
SE2	1.97%	**5.80%**	2.47%	2.36%
SE3	2.78%	1.70%	**3.71%**	0%
SE13	**4.38%**	-	-	-
SE15	3.39%	**4.78%**	3.75%	2.50%
ALL	3.26%	**4.30%**	3.14%	2.31%

Table 7. Percentage of Parts of Speech of Newly Mistaken Word Senses in Each Evaluation Set (BEM2)

To analyze in detail the findings presented in Table 5, we investigated the percentage of newly mistaken word senses parts of speech in the ALL evaluation set when using examples of word senses. The results of the survey are shown in Table 6. The survey results show that the use of example word senses increases the number of mistakes most frequently in the identification of verb senses. Therefore, we consider that to improve the accuracy of the WSD, it is necessary to reconsider the method of extracting examples of word senses related to verbs.

To analyze the cause of the decrease in accuracy in the evaluation sets other than senseval3, we examined the proportion of parts of speech of newly mistaken word senses in BEM2 in each evaluation set. The results of the survey are shown in Table 7. The survey results show that the senseval3 evaluation set with improved accuracy has fewer errors in verb senses, while the other evaluation sets have more errors in verb senses. This suggests that the accuracy of identifying verb senses contributes to the difference in accuracy of each evaluation set. One possible reason for the high number of errors in verb senses could be that only examples of word senses containing words of the same type as the target word were used. Therefore, we consider that accuracy could be improved by increasing the number of examples of word senses used by extracting examples of word senses that include words converted to the past tense, plural, etc.

7 Conclusion and Future Work

In this study, we analyzed the effectiveness of examples of word senses in WSD by using examples of word senses retrieved from WordNet 3.0 as sense embedding representations and incorporating them into BEM. The results showed that the use of examples of word senses decreased the overall performance, but the model using the target word vector of the examples of word senses slightly improved the F1 score on the Senseval-3 evaluation set. Additionally, the F1 score of zero-shot words was improved. Thus, we expect that although examples of word senses have a negative impact on BEM, they can be effective if examples of word senses are used in a different way.

For future work, we are considering reexamining the extraction method when using the target word vector of examples of word senses, such as targeting not only those that are isomorphic to the target word, but also those that have been transformed into plural or past tense forms. We are also considering other ways to use examples of word senses to mitigate data bias, such as using examples of word senses only in the case of LFS without using examples of word senses in the case of MFS.

References

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Word sense disambiguation: A unified evaluation framework and empirical comparison. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 99–110.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 288–297.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viegas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. arXiv preprint arXiv:1906.02715.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5300–5309.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. Incorporating glosses into neural word sense disambiguation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2473–2482, Melbourne, Australia. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1402–1411, Brussels, Belgium. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In Proceedings of the workshop on Human Language Technology, pages 303–308. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Sackinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In Advances in neural information processing systems, pages 737–744.

Joyce Yue Chai and Alan W Biermann. 1999. The use of word sense disambiguation in an information extraction system. In Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI '99/IAAI '99), pages 850-855.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, pages 21 – 24.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, pages 24–26. ACM.

Nazreena Rahman and Bhogeswar Borah. 2020. Improvement of query-based text summarization using word sense disambiguation. Complex Intelligent Systems, vol. 6 No.7, pages 75-85.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1591–1600.

Quang-Phuoc Nguyen, Anh-Dung Vo, Joon-Choul Shin and Cheol-Young Ock. 2018. Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean. in IEEE Access, vol. 6, pages 38512-38523.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 222–231.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007), pages 87–92.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In IJCAI.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1006-1017.

Effective Use of Japanese Dictionary Definition Sentences in Learning Hierarchical Embedding of Dictionaries

Yuki Ishii
Ibaraki University
22nm702y@vc.ibaraki.ac.jp

Minoru Sasaki
Ibaraki University
minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract

Existing knowledge graph-based learning methods for word sense disambiguation use word-to-word relations to learn models, but do not learn using the hierarchical relations of word senses in a single word. In addition, even when defining sentences in a Japanese dictionary are applied, the accuracy of judging the hierarchical relationship of word senses is poor, and the effect of knowledge graph embedding learning is not fully achieved. This study analyzes how to edit dictionary descriptions to improve the accuracy of models that judge the hierarchical relationship between senses in a Japanese dictionary. The results of the analysis showed that the accuracy of the unedited dictionary was 60.9%, while the accuracy of the edited dictionary was 83.3%, confirming the improved performance of the model.

1 Introduction

Word sense disambiguation (WSD) is the task of identifying which sense is used in a sentence for a polysemous word in the sentence. In the last few years, many knowledge graph-based approaches have been studied that do not require the cost of word sense labeling. Related works include GlossBERT (Huang et al., 2019) and BEM (Blevins and Zettlemoyer, 2020), which combine word sense definition sentences with supervised learning, and EWISE (Kumar et al., 2019) and EWISER (Bevilacqua and Navigli, 2020), which use word-to-word relations in knowledge data. However, these methods do not learn models using the hierarchical relationship of word senses in words. When the EWISE system learned the hierarchical relationship of word senses using the Japanese dictionary definition sentences in their original form, it was found that the relationship judgment accuracy of the generated models was low, and the knowledge information was not sufficiently learned. In the Iwanami Japanese Dictionary, there are definition sentences that only describe usage expressions and part-of-speech expressions such as "《attached to noun》" and "((adjective))". In Japanese, it is expressed as "《名詞に付けて》" and "((形))". Such definitions are likely to exist in multiple dictionaries and overlap in content, thus inhibiting learning of the hierarchical relationship between senses. Therefore, in addition to the existing relations between words, we will learn the hierarchical relations of word senses and analyze how the contents of the Japanese dictionary can be edited and modified to improve the accuracy of the hierarchical relation judgment model of word senses. Learning a model for judging the hierarchical relationship of word senses is an initial task to improve the performance of the knowledge graph embedding system.

To investigate how well the information in the knowledge graph can be embedded, we use the knowledge graph embedding learning of the EWISE system, which provides accuracy in judging the relationship between triples of data in the knowledge data.

2 Related Work

Many approaches based on knowledge graphs have been studied.

In GlossBERT, a pre-trained BERT encoder is given a context and a word definition sentence, and is trained to judge whether the word definition sentence correctly describes the usage of the target word. In BEM, two encoders are used for the GlossBERT approach to learn the distributed representations of the context and the word definition sentences separately. These studies combine word definition sentences with supervised learning.

In EWISE, distributed representations of word senses are learned from WordNet word-word relations and incorporated into the WSD task. In EWISER, the weights given to words are learned using WordNet word-to-word relationships, and the resulting matrices are incorporated into the WSD task. These studies use word-to-word relationships in knowledge data.

However, in these related works, no model learning has been conducted using the hierarchical relationship of word senses in a single word. In this study, the model is trained using a single word hierarchical relationship.

3 Hierarchical relation judgment system

The knowledge graph embedding system of the EWISE system is used to judgment the hierarchical relationship between senses in a Japanese dictionary. The system is shown in Figure 1.

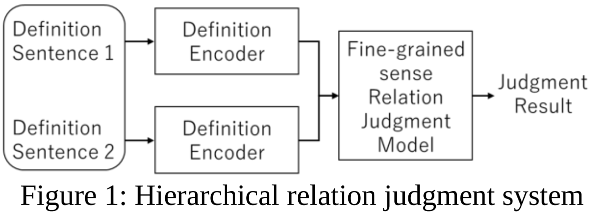


Figure 1: Hierarchical relation judgment system

3.1 Definition Sentence Encoder

The definition sentence encoder uses the BiLSTM Max encoder (Conneau et al, 2017). The fixed-length representation obtained by inputting definition sentences to BiLSTM and Max Pooling the output is the output of the definition sentence encoder.

3.2 Fine-grained Sense Relation Judgment Model

The fine-grained sense relation judgment model takes as input a distributed representation of two Japanese definitions and judges the hierarchical relation between the definitions. ConvE (Dettmers et al., 2013) is used for this fine-grained sense relation judgment model. A knowledge graph usually consists of a set K of N triples (h, l, t) consisting of two entities (h, t) and one relation (l). h is the head entity and t is the tail entity. ConvE formulates the scoring function $\psi_l(e_h, e_t)$ for a triple (h, l, t) as:

$$\psi_l(e_h, e_t) = f\big(\mathrm{vec}\big(f([\overline{e_h}; \overline{e_l}] * w)\big)W\big)e_t \qquad (1)$$

where e_h and e_t are the parameters of the entity, e_l is the parameter of the relation, $\bar{x}$ is a 2-dimensional deformation of x, w is a 2-dimensional convolution filter, $\mathrm{vec}(x)$ is a vectorization of x, w is a linear transformation, and f is a normalized linear unit. For the target head entity h, we compute the score $\psi_l(e_h, e_t)$for each entity in the graph as a tail entity. Probability estimates for the validity of a triple (h, l, t) are obtained by applying a sigmoid function to the scores:

$$p = \sigma\big(\psi_l(e_h, e_t)\big) \qquad (2)$$

3.3 Model Learning

Figure 2 shows the training flow of the definition sentence encoder and the fine-grained sense relation judgment model. The training data are knowledge data describing the hierarchical relationship of word meanings and definition sentences associated with the definition sentence IDs. The definition sentence is decomposed into morphemes, and the distributed representation of each morpheme is input to the definition sentence encoder. The distributed representation of morphemes is pre-trained using fastText (Bojanowski et al., 2016) and GloVe (Pennington et al., 2014). The word sense relations are transformed into embedding vectors by the embedding layer in the inter-sense relation judgment model. The parameters of the fine-grained sense relation judgment model are initially set to the parameters of the initial model trained with only triples (h, l, t) of word sense hierarchical relations. Equation (1) is modified to Equation (3) to learn the definition sentence encoder.

$$\psi_l(e_h, e_t) = f\left(\text{vec}\left(f([\,\overline{q(h)}; \overline{e_l}\,] * w)\right) W\right) e_t \quad (3)$$

where $q(.)$ is the definition sentence encoder and the head entity is the encoded definition of the entity. The parameters of the model are updated based on the estimates p and the labels. The loss function uses the binary cross-entropy:

$$L_C = -\frac{1}{N}\sum_i (t_i \cdot log(p_i) + (1 - t_i) \cdot log(1 - p_i)) \quad (4)$$

where t_i is 1 if the triple (h, l, t) is appropriate, 0 otherwise. p_i is an estimate of the score shown in Equation (3).

Section 5.1 shows the training data, the distributed representation of morphemes, and Section 5.3 details the evaluation method of the model.

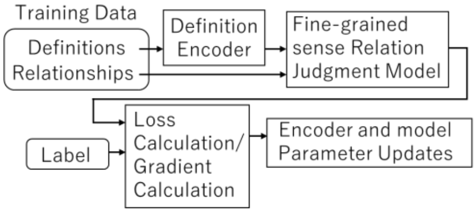

Training Data
Definitions
Relationships
Definition Encoder
Fine-grained sense Relation Judgment Model
Label
Loss Calculation/ Gradient Calculation
Encoder and model Parameter Updates

Figure 2: Model Learning Flow

4 Editing and modifying the definition text

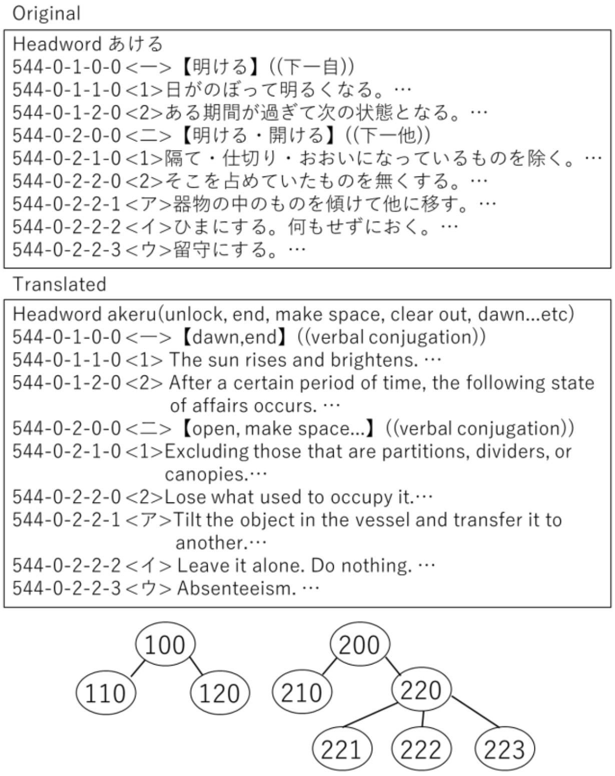
4.1 Editing Definition Sentences

Edit the Iwanami Japanese Dictionary to correspond to the WordNet (Miller, 1995) triple data used in EWISE.

4.1.1 Assigning IDs to Definition Sentences

The Iwanami Japanese Dictionary's definition sentence IDs are assigned by combining the headword number (1 to 5 digits), major category number (1 digit), middle category number (1 digit), and minor category number (1 to 2 digits) in this order, excluding the compound word number.

The hierarchical relationship of word senses is learned on the two relationships of hypernym and hyponym from the classification method of the Iwanami Japanese Dictionary. The hierarchical relationship is a tree structure as shown in Figure 3. The parent of a tree structure is a hypernym of its children, and its children are hyponym of the parent. When recording triples, only the relationships between nodes connected by edges are recorded.


Original

Headword あける
544-0-1-0-0<一>【明ける】((下一自))
544-0-1-1-0<1>日がのぼって明るくなる。…
544-0-1-2-0<2>ある期間が過ぎて次の状態となる。…
544-0-2-0-0<二>【明ける・開ける】((下一他))
544-0-2-1-0<1>隔て・仕切り・おおいになっているものを除く。…
544-0-2-2-0<2>そこを占めていたものを無くする。…
544-0-2-2-1<ア>器物の中のものを傾けて他に移す。…
544-0-2-2-2<イ>ひまにする。何もせずにおく。…
544-0-2-2-3<ウ>留守にする。…

Translated

Headword akeru(unlock, end, make space, clear out, dawn...etc)
544-0-1-0-0<一>【dawn,end】((verbal conjugation))
544-0-1-1-0<1> The sun rises and brightens. …
544-0-1-2-0<2> After a certain period of time, the following state of affairs occurs. …
544-0-2-0-0<二>【open, make space...】((verbal conjugation))
544-0-2-1-0<1>Excluding those that are partitions, dividers, or canopies.…
544-0-2-2-0<2>Lose what used to occupy it.…
544-0-2-2-1<ア>Tilt the object in the vessel and transfer it to another.…
544-0-2-2-2<イ> Leave it alone. Do nothing. …
544-0-2-2-3<ウ> Absenteeism. …

Figure 3: Hierarchical relationship of word senses (The number in the node corresponds to the right three digits of the dictionary ID.)

4.2 Change Definition Sentence

To improve the accuracy of judging the relationship between triples of knowledge data, we propose the following settings A, B, ① to ⑤.

4.2.1 Setting A and B

There are word definitions in which the definition is written only with the double parentheses "(( ))", double mountain brackets "《》", and turtle-shell bracket "[ ]", which do not explain the meaning of the word, so the definition itself is deleted. Deleting a definition sentence affects the total number of triples of knowledge data differently from changes in subsequent clauses. Experiments were conducted for setting A, in which the Iwanami Japanese Dictionary was not edited, and setting B, in which definition sentences that do not explain word meanings were deleted, and the setting with the highest accuracy was carried over to the changes in subsequent clauses.

4.2.2 Setting ① and ①*.

The three types of expressions, double parentheses "(( ))", double parentheses "《》", and parentheses "< >", indicate the attributes, uses, and classification of the headword, and were deleted from the definition text. For the turtle-shell bracket "[ ]" expression, a similar expression exists in the WordNet definition text, but the setting ①* is deleted to check whether the category name is necessary. Symbols were removed and replaced because they were superfluous expressions. The symbols deleted were "↓", "△", and "×", and the symbols replaced were "▽" with"。" and "【】" with "『』". In addition, full-width alphabetic characters have been changed to half-width characters. Example sentences are described in the dictionary, but the headwords included in the example sentences are replaced by hyphens "-". The hyphen "-" is replaced by the headword (hiragana).

4.2.3 Setting ②

In the case of hiragana, it is preferable to convert the words to kanji to obtain an accurate representation of the variance because of the presence of words with homonymous meanings. Since many of the headwords have more than one Kanji character, those that can be replaced one-to-one were replaced with Kanji characters.

4.2.4 Setting ③

There are many expressions in the definitions that cannot be the direct meaning of a word. The following is a list of those that were deleted.

- The sentences following "▽" provide explanations that go beyond the meaning of the target word.
- The sentence followed by "派生｜" ("Derivatives|") indicates the derivation of the declension.
- To remove reading kana, delete the content enclosed in full-width parentheses "( )" if it is in hiragana only.
- The content enclosed by single square brackets "( )" indicates the figure number and annotation number in the dictionary.

4.2.5 Setting ④

In setting ④, delete the contents enclosed in full-width parentheses "( )" regardless of the contents.

4.2.6 Setting ⑤

In the Iwanami Japanese Dictionary, some example sentences surrounded by "「」" are sentences without headwords. In such cases, it is difficult to complete the sentence and incomplete example sentences remain. Therefore, if a single sentence delimited by a punctuation mark is surrounded by key brackets "「」" in the definition sentence, the entire example sentence is deleted.

5 Experiment

5.1 Experimental Data

- The hierarchical relationship of word senses
  Triples of training data from the Iwanami Japanese Dictionary, in which the IDs of definitions (h, t) and the hierarchical relationship of meanings (l) are recorded in the order h, l, t, are used. Setting A with no editing obtained 24040 triples, while setting B, which removed definition sentences without word descriptions, obtained 10842 triples. The triples of training data obtained are divided into training, development, and test data in the ratio of 8:1:1.
- Definition sentences
  Avoiding duplicate definition sentences may have improved accuracy. The definition sentences of the word senses associated with the definition sentence IDs of the knowledge data are used as experimental data. MeCab and mecab-ipadic-NEologd were used to separate words in the definition sentences. The distributed representation of morphemes in the definition sentences is based on a corpus of dumped data from the full-page articles of Japanese Wikipedia updated on October 10, 2021, which was pre-trained using fastText and GloVe.

Table 1: Experimental results
(Total vocabulary (total number of words in the definition sentence), Words Vector Number (number of words for which a distributed representation was obtained), Ratio (number of vocabulary vectors/total vocabulary))

Setting/ Result Detail	fastText				GloVe			
	Words Vector Number	Total vocabulary	Ratio	MRR	Words Vector Number	Total vocabulary	Ratio	MRR
A	60067	72619	0.8272	0.60853	60066	72618	0.8272	0.60957
B	60147	72752	0.8267	0.82286	60146	72751	0.8267	0.82203
B①*	66813	81039	0.8245	**0.83420**	66812	81038	0.8245	0.82640
B①②	68311	79958	0.8543	0.82890	68310	79957	0.8543	0.82590
B①②③	62275	71860	0.8666	0.83137	62274	71859	0.8666	0.82874
B①②③④	61449	70805	0.8679	0.83177	61449	70805	0.8679	0.82928
B①②③④⑤	43267	48612	0.8900	0.82251	43266	48611	0.8900	0.82098
B①*③④	60417	72181	0.8370	0.83341	60416	72180	0.8370	**0.83135**

5.2 Hyperparameter

Below are the hyperparameters for each study, modified from the default values.

When training the definition encoder and the fine-grained sense relation judgment model, the batch size was changed to 64, and the number of epochs was set to 160. The number of epochs for the initial model was set to 300, while for GloVe pre-training, the number of surrounding words considered for training was set to 10, and the dimensionality of the distributed representation was set to 300.

5.3 Evaluation Method

The MRR (Mean Reciprocal Rank) is used as the evaluation method for the fine-grained sense relation judgment model. MRR is one of the ranking evaluation indexes, which evaluates a model with a target of the relationship estimates output from the fine-grained sense relation judgment model. In this case, MRR is given as:

$$MRR = \frac{1}{|U|}\sum_{u \in U}\frac{1}{k_u} \qquad (5)$$

where u is the target triple, U is the total triple, and k_u is the order in which the entities whose relations were correctly judged for the target triple appeared.

6 Results

Table 1 shows the experimental results. Deletion of definitions that do not explain word meanings was effective, while deletion of expressions in the dictionary and deletion of additional information slightly improved the accuracy of judgments.

The accuracy improvement was significantly greater when definition sentences that did not explain the word meanings were removed. Most of these definitions are conjugational, which means that there are multiple definitions that are identical word-for-word.

Changing the definition sentence affects the number of words in the entire definition sentence and the number of words from which the initial value of the distributed expression can be obtained. Regardless of the number of word vectors or the number of words in the vocabulary, it is acceptable to remove expressions that are deemed unnecessary to explain the meaning of the target word.

In setting ⑤, it is difficult to remove example sentences that cannot be complemented, which may have reduced the accuracy of the results.

In addition, training with fastText showed a slight improvement in accuracy compared to GloVe. Distributed representations of words learned using subword information were shown to improve decision accuracy compared to using global co-occurrence information.

7 Conclusion and Future Work

Experimental results showed that learning by removing or changing expressions that do not explain the meaning of a definition sentence is beneficial for improving the performance of the hierarchical relation judgment model. The hierarchical relationship judgment model obtained in this study is expected to improve the performance of knowledge graph embedding systems. In the future, we plan to study the effectiveness of this method by conducting word sense disambiguation experiments using the knowledge graph embedding system.

References

Huang, L., Sun, C., Qiu, X. and Huang, X. "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3509-3514, 2019.

Blevins, T. and Zettlemoyer, L. "Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders", In Proceedings of the 58th Association for Computational Linguistics (ACL2020), pp. 1006-1017, 2020.

Sawan, K. Sharmistha. J. Karan, S. and Partha T."Zero-shot word sense disambiguation using sense definition embeddings." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5670-5681, 2019.

Bevilacqua M. and Navigli R. "Breaking Through the (80%) Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020), pp. 2854-2864, 2020.

Alexis, C. Douwe, K. Holger S. Loïc, B. and Antoine, B. "Supervised learning of universal sentence representations from natural language inference data." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics, 2017.

Tim, D. Pasquale, M. Pontus, S. and Sebastian, R. "Convolutional 2d knowledge graph embeddings." In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

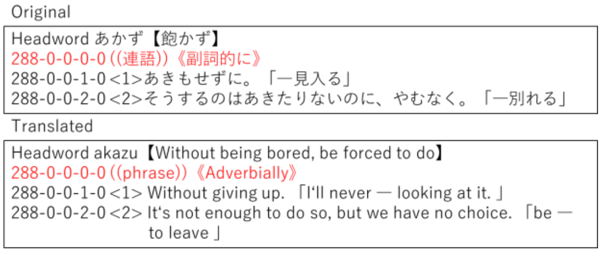
Piotr, B. Edouard, G. Armand, J. and Tomas, M. "Enriching Word Vectors with Subword Information." In Transactions of the Association for Computational Linguistics, Volume 5, pp. 135–146, 2016.

Jeffrey, P. Richard, S. and Christopher, M. "GloVe: Global Vectors for Word Representation." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.

George, M. "Wordnet: a lexical database for english." Communications of the ACM, 38(11):39– 41, 1995.

Appendix: Specific examples of the settings

Specific examples of the settings shown in section 4.2 of the main text are given below. These definitions are taken from the Iwanami Japanese Dictionary, 5th edition.



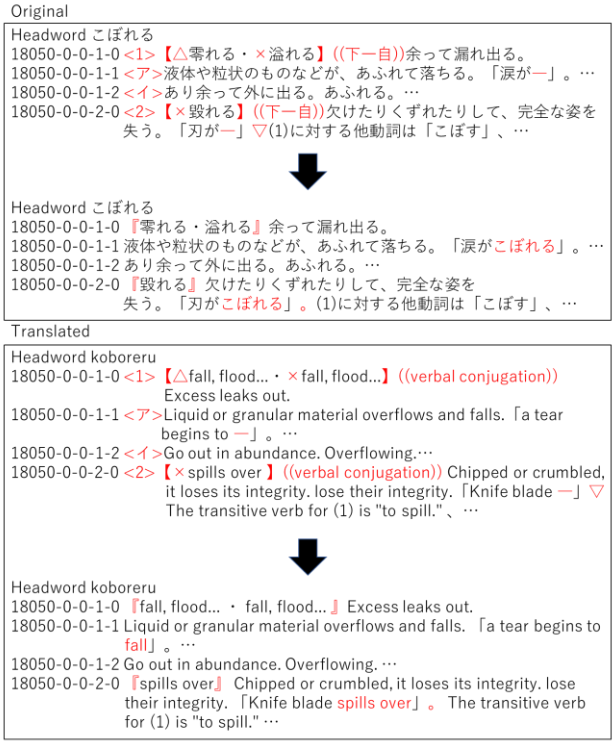
Original

Headword あかず【飽かず】
288-0-0-0-0 ((連語))《副詞的に》
288-0-0-1-0 <1>あきもせずに。「一見入る」
288-0-0-2-0 <2>そうするのはあきたりないのに、やむなく。「一別れる」

Translated

Headword akazu【Without being bored, be forced to do】
288-0-0-0-0 ((phrase))《Adverbially》
288-0-0-1-0 <1> Without giving up. 「I'll never — looking at it. 」
288-0-0-2-0 <2> It's not enough to do so, but we have no choice. 「be — to leave 」

Figure A: Example of specific changes to setting B (removing red text)



Original

Headword こぼれる
18050-0-0-1-0 <1> 【△零れる・×溢れる】((下一自))余って漏れ出る。
18050-0-0-1-1 <ア>液体や粒状のものなどが、あふれて落ちる。「涙が一」。…
18050-0-0-1-2 <イ>あり余って外に出る。あふれる。…
18050-0-0-2-0 <2> 【×毀れる】((下一自))欠けたりくずれたりして、完全な姿を失う。「刃が一」▽(1)に対する他動詞は「こぼす」、…

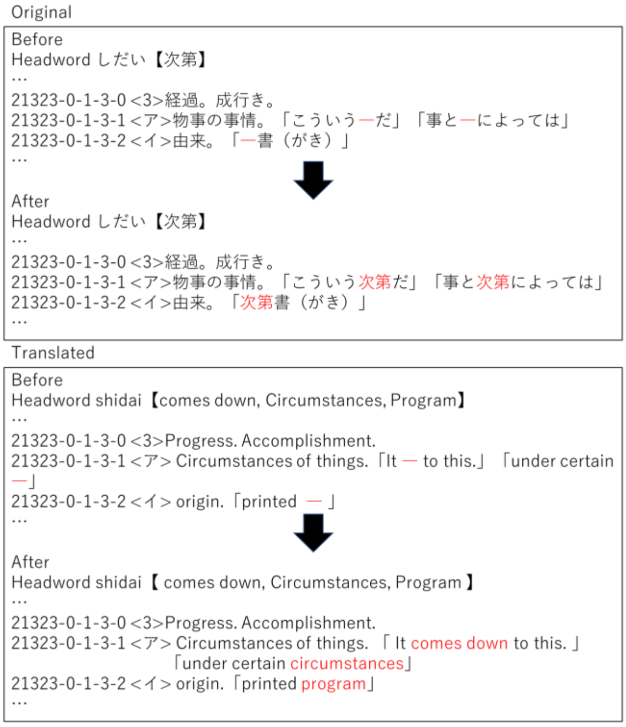
Headword こぼれる
18050-0-0-1-0 『零れる・溢れる』余って漏れ出る。
18050-0-0-1-1 液体や粒状のものなどが、あふれて落ちる。「涙がこぼれる」。…
18050-0-0-1-2 あり余って外に出る。あふれる。…
18050-0-0-2-0 『毀れる』欠けたりくずれたりして、完全な姿を失う。「刃がこぼれる」。(1)に対する他動詞は「こぼす」、…

Translated

Headword koboreru
18050-0-0-1-0 <1> 【△fall, flood... ・ ×fall, flood...】((verbal conjugation)) Excess leaks out.
18050-0-0-1-1 <ア>Liquid or granular material overflows and falls. 「a tear begins to —」。…
18050-0-0-1-2 <イ>Go out in abundance. Overflowing.…
18050-0-0-2-0 <2> 【×spills over 】((verbal conjugation)) Chipped or crumbled, it loses its integrity. lose their integrity. 「Knife blade —」▽ The transitive verb for (1) is "to spill." 、…

Headword koboreru
18050-0-0-1-0 『fall, flood... ・ fall, flood... 』Excess leaks out.
18050-0-0-1-1 Liquid or granular material overflows and falls. 「a tear begins to fall」。…
18050-0-0-1-2 Go out in abundance. Overflowing. …
18050-0-0-2-0 『spills over』 Chipped or crumbled, it loses its integrity. lose their integrity. 「Knife blade spills over」。 The transitive verb for (1) is "to spill." …

Figure B: Examples of specific changes to setting ①



Original

Before
Headword しだい【次第】
…
21323-0-1-3-0 <3>経過。成行き。
21323-0-1-3-1 <ア>物事の事情。「こういう一だ」「事と一によっては」
21323-0-1-3-2 <イ>由来。「一書（がき）」
…

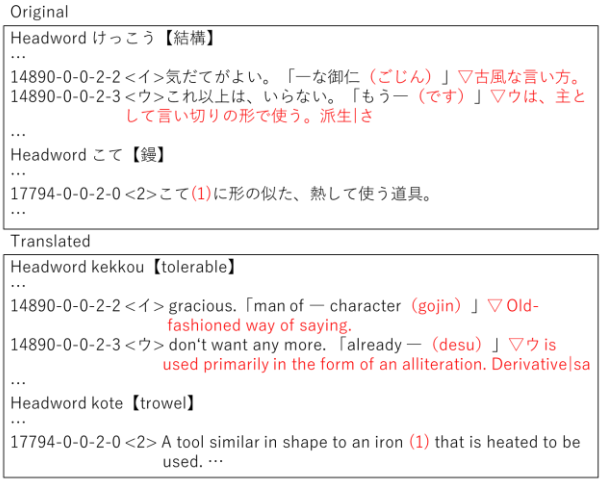
After
Headword しだい【次第】
…
21323-0-1-3-0 <3>経過。成行き。
21323-0-1-3-1 <ア>物事の事情。「こういう次第だ」「事と次第によっては」
21323-0-1-3-2 <イ>由来。「次第書（がき）」
…

Translated

Before
Headword shidai【comes down, Circumstances, Program】
…
21323-0-1-3-0 <3>Progress. Accomplishment.
21323-0-1-3-1 <ア> Circumstances of things. 「It — to this.」「under certain —」
21323-0-1-3-2 <イ> origin. 「printed — 」
…

After
Headword shidai【 comes down, Circumstances, Program 】
…
21323-0-1-3-0 <3>Progress. Accomplishment.
21323-0-1-3-1 <ア> Circumstances of things. 「 It comes down to this. 」「under certain circumstances」
21323-0-1-3-2 <イ> origin. 「printed program」
…

Figure C: Specific examples of changes in setting ②



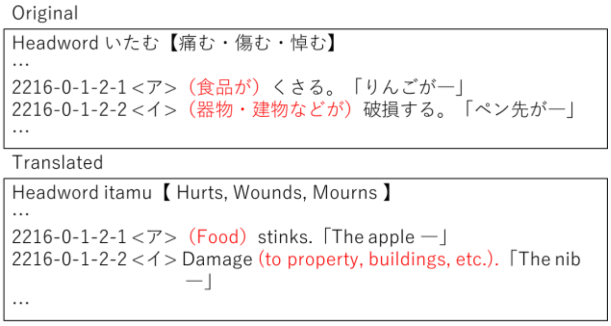
Original

Headword けっこう【結構】
…
14890-0-0-2-2 <イ>気だてがよい。「一な御仁（ごじん）」▽古風な言い方。
14890-0-0-2-3 <ウ>これ以上は、いらない。「もう一（です）」▽ウは、主として言い切りの形で使う。派生|さ
…
Headword こて【鏝】
…
17794-0-0-2-0 <2>こて(1)に形の似た、熱して使う道具。
…

Translated

Headword kekkou【tolerable】
…
14890-0-0-2-2 <イ> gracious. 「man of — character（gojin）」▽ Old-fashioned way of saying.
14890-0-0-2-3 <ウ> don't want any more. 「already —（desu）」▽ウ is used primarily in the form of an alliteration. Derivative|sa
…
Headword kote【trowel】
…
17794-0-0-2-0 <2> A tool similar in shape to an iron (1) that is heated to be used. …

Figure D: Example of specific change in setting ③ (removing red text)



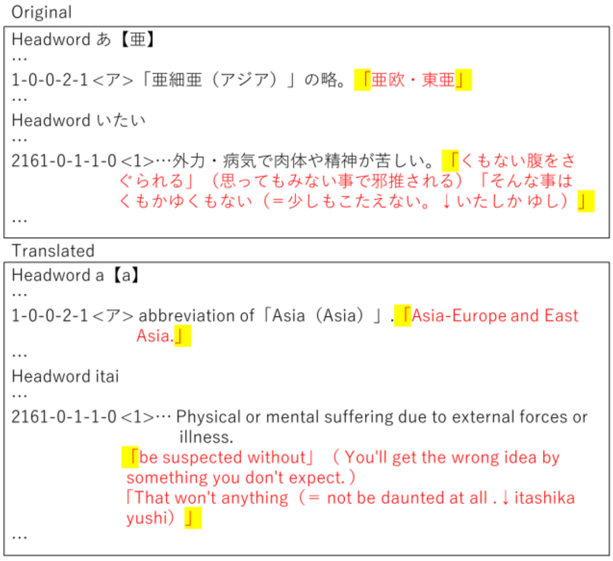
Original

Headword いたむ【痛む・傷む・悼む】
…
2216-0-1-2-1＜ア＞（食品が）くさる。「りんごが—」
2216-0-1-2-2＜イ＞（器物・建物などが）破損する。「ペン先が—」
…

Translated

Headword itamu【 Hurts, Wounds, Mourns 】
…
2216-0-1-2-1＜ア＞（Food）stinks.「The apple —」
2216-0-1-2-2＜イ＞Damage (to property, buildings, etc.).「The nib —」
…

Figure E: Example of a specific change in setting ④ (removing red text)



Original

Headword あ【亜】
…
1-0-0-2-1＜ア＞「亜細亜（アジア）」の略。「亜欧・東亜」
…
Headword いたい
…
2161-0-1-1-0＜1＞…外力・病気で肉体や精神が苦しい。「くもない腹をさぐられる」（思ってもみない事で邪推される）「そんな事はくもかゆくもない（＝少しもこたえない。↓いたしか ゆし）」
…

Translated

Headword a【a】
…
1-0-0-2-1＜ア＞abbreviation of「Asia（Asia）」.「Asia-Europe and East Asia.」
…
Headword itai
…
2161-0-1-1-0＜1＞… Physical or mental suffering due to external forces or illness.
「be suspected without」（You'll get the wrong idea by something you don't expect.）
「That won't anything（= not be daunted at all . ↓ itashika yushi）」
…

Figure F: Example of a specific change in setting ⑤ (removing red text)

The Wikipedia corpus used for pre-training of word distributed representations was obtained from

https://ja.wikipedia.org/wiki/Wikipedia:%E3%83%87%E3%83%BC%E3%82%BF%E3%83%99%E3%83%BC%E3%82%B9%E3%83%80%E3%82%A6%E3%83%B3%E3%83%AD%E3%83%BC%E3%83%89

Bi-directional Cross-Attention Network on Vietnamese Visual Question Answering

Duy-Minh Nguyen-Tran
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
20c11041@student.hcmus.edu.vn

Tung Le
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
lttung@fit.hcmus.edu.vn

Minh Le Nguyen
Japan Advanced Institute of Science
and Technology, Nomi, Ishikawa, Japan
nguyenml@jaist.ac.jp

Huy Tien Nguyen ✉
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
ntienhuy@fit.hcmus.edu.vn

Corresponding author: Huy Tien Nguyen
ntienhuy@fit.hcmus.edu.vn

Abstract

Visual Question Answering (VQA) has arisen in recent public interest thanks to its applicability in many different fields. However, it requires understanding the combination of pictures and questions, which is highly challenging in both vision and language processing. Many previous works have achieved remarkable results to address this problem in many different languages. However, in the Vietnamese language, the VQA problem has not made significant progress due to the lack of data and fundamental systems. Therefore, we propose a model specifically designed and optimized for the Vietnamese Visual Question Answering problem. Our model leverages the strength of pre-trained models as well as presents Bi-directional Cross-attention architecture to learn visual and textual features more effectively. Through experimental results and ablation studies, the proposed approach obtains promising results against the existing models for Vietnamese on the ViVQA dataset.

1 Introduction

Together with the remarkable development in Computer Vision and Natural Language Processing, many problems requiring the combination of both images and languages were raised such as Image Captioning (Wang et al., 2022; Hu et al., 2021), Visual Question Answering (VQA) (Le et al., 2021a; Le et al., 2020), Visual Question Classification (Le. et al., 2022), etc. Among them, VQA is a potential area receiving a lot of attention in both research and industry. Furthermore, the recent approaches in VQA also achieved promising results. The goal of most VQA systems is to digest the content of a given image and answer the related questions. Compared to other tasks like Image Captioning and Visual Question Classification, VQA requires a deeper understanding of visual and textual inputs to give appropriate answers. In practice, VQA can be applied to various scenarios such as human-machine interaction, medical assistance, and automatic customer service or recommendation.

Undoubtedly, the VQA model plays an important role in the cutting edge of multi-modal approaches between images and languages. In recent years, many VQA approaches have been introduced and achieved promising results such as LXMERT (Tan and Bansal, 2019), SIMVL (Wang et al., 2021), OSCAR (Li et al., 2020). However, current VQA systems are almost addressed in English, Japanese, Chinese, and a few other languages. However, VQA for Vietnamese has not been strongly developed due to data limitations. Therefore, we introduce an optimized and fine-tuned VQA model specifically for the Vietnamese language. Besides, we propose bi-directional cross-attention architecture by adjusting the multi-head Attention in Transformer (Vaswani et al., 2017) to optimize learning image-question pairs with the Vietnamese language. To demonstrate the effectiveness of the method, our model is evaluated on the ViVQA dataset (Tran et al., 2021) which is built and adjusted based on the characteristics of the Vietnamese language. Through experiments, our model proves its efficiency and outperforms the competitive baselines.

The major contributions of this paper are concluded as follows: (i) We deploy to use the Vision - Text Transformer model to optimize the feature extraction of vision and language for Vietnamese. (ii) We propose a bi-directional cross-attention architecture by adjusting the attention structure of the transformer. Our proposed component is efficient to learn the combination and relationship between visual and linguistic features for Vietnamese. (iii) Through experiments and ablation studies, our proposed model achieves superior results compared to the existing approaches in the VQA dataset for the Vietnamese language.

2 Related Works

In the early stages of the Visual Question Answering problem, previous architectures are often built by two independent networks in CV and NLP to understand image and text features combined by the external components such as vector operations (Le et al., 2021b) and stacked attention (Le et al., 2021a). In particular, typical approaches in traditional VQA systems used Convolution Neural Networks (CNNs) for image embedding while a question was embedded by Recurrent Neural Networks (RNNs) (Goyal et al., 2017). Then, visual and textual features are combined by an attention mechanism.

In recent years, transfer learning has arisen the community's interest in many works to take advantage of huge datasets via self-supervised learning. There are more and more pre-trained models, which are the cornerstone for many areas such as BERT (Devlin et al., 2018), Vision Transformer (Dosovitskiy et al., 2021), and Speech-BERT (Chuang et al., 2020), etc. In recent VQA approaches, pre-trained models are also applied in many feature extraction modules to inherit the development of the CV and NLP domain. By using pre-trained models that have been trained on huge datasets, the feature extraction components are more efficient thereby improving the efficiency of the model.

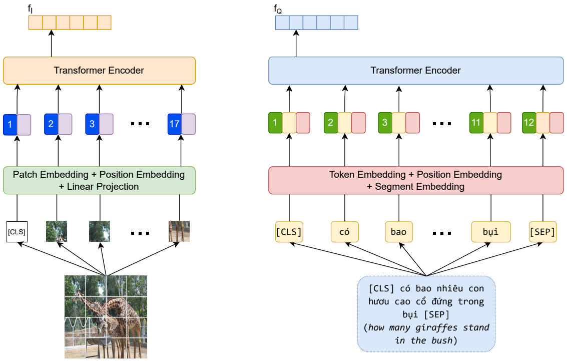
Together with the necessity of feature extraction, the combination of RNNs and CNNs is considered one of the most important components in VQA systems. In understanding signals from images and texts, in recent years, with the development of deep learning, the Transformer model has been introduced and achieved many admirable results in both Natural Language Processing and Computer Vision fields. Transformer's Attention Architecture was applied to many VQA systems and achieved promising results. LXMERT (Tan and Bansal, 2019), one of the best models for the VQA problem, uses the self-attention and multi-head attachment architecture of Transformer to propose a cross-modality encoder architecture in combining visual and textual features. In addition, another famous model, SIMVLM (Wang et al., 2021), applied Transformer encoder architecture in feature extraction and achieved superior results against previous models in the VQA problem.

However, the Visual Question Answering problem for the Vietnamese language (ViVQA) developed at a modest rate due to the lack of resources. Among many previous works in ViVQA, Tran et al. (Tran et al., 2021) use PhoW2Vec (Nguyen and Nguyen, 2020) to address Vietnamese questions and Hierarchical Co-Attention (Lu et al., 2016) to combine visual and textual features. This model currently achieves the best performance in the ViVQA dataset. However, the performance of the model is still very limited against VQA systems in other languages. In addition, due to specific language characteristics, recent approaches in English also meet the difficulties as being applied to the Vietnamese dataset. This issue inspired us to build a VQA model specifically for Vietnamese. By using pre-trained models and the proposed bi-directional cross-attention architecture, our model improves the efficiency compared to previous models by deploying a feature understanding module compatible with the Vietnamese language.

3 Model Architecture

3.1 Visual and Textual Feature Extraction

Our model leverages the power of pre-trained models in the feature extraction of both images and texts. Using pre-trained models makes feature extraction more effective with prior knowledge from huge datasets in both CV and NLP domains. Details of our feature extraction modules are shown in Figure 1.



f_I
Transformer Encoder
1
2
3
...
17
Patch Embedding + Position Embedding + Linear Projection
[CLS]
...
f_Q
Transformer Encoder
1
2
3
...
11
12
Token Embedding + Position Embedding + Segment Embedding
[CLS]
có
bao
...
bụi
[SEP]
[CLS] có bao nhiêu con hươu cao cổ đứng trong bụi [SEP]
(*how many giraffes stand in the bush*)

Figure 1: Detail of Feature Extraction Modules: Images are embedded by the pre-trained Vision Transformer model (left) and Questions are embedded by the pre-trained PhoBERT model (right)

In particular, images are embedded by the pre-trained Vision Transformer model (Dosovitskiy et al., 2021). With the success of the Transformer model in the field of Natural Language Processing (NLP), Alexey Dosovitskiy et al. (Dosovitskiy et al., 2021) applied this architecture to Computer Vision and achieved better results than the state-of-the-art models in image classification. The Vision Transformer model divides an image into a sequence of regions called patches and treats them as words in a sentence. In our image embedding, a special patch $[CLS]$ is added to the first position of the visual patches to represent the aggregated information of an image. After being processed by Transformer architecture, visual features f_I are extracted from the $[CLS]$ representation vector.

For question embedding, a pre-trained PhoBERT model (Nguyen and Nguyen, 2020) is used to extract question features in our model. Although the pre-trained model BERT (Devlin et al., 2018) is considered the best performing model in many NLP tasks, it has not achieved good results when applied to Vietnamese datasets. Therefore, based on BERT, PhoBert is built and trained on the Vietnamese dataset to address the typical characteristics of this language. In particular, for extracting textual features, we put two special characters $[CLS]$ and $[SEP]$ in the first and last positions respectively. After going through pre-trained PhoBERT, we utilize the outputs of $[CLS]$ vector as question features f_Q.

3.2 Bi-directional Cross-Attention Network

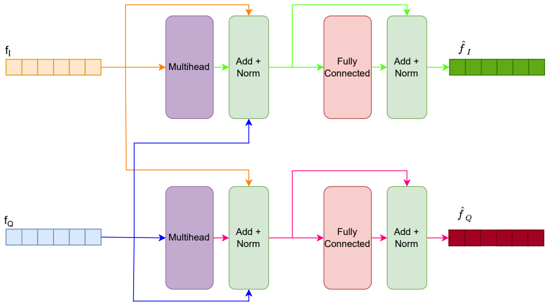


Figure 2: Bi-directional Cross-Attention

To understand visual and linguistic combinations and relationships, we propose a Bi-directional Cross-Attention architecture for simultaneous learning of textual and visual features. Bi-directional Cross-Attention is inspired by Transformer's Attention with some adjustments. While Transformer's Attention is only suitable for paying attention to individual features, our proposed architecture is capable of learning simultaneous and mutual relationships between images and languages. Our proposed component is visualized in Figure 2. Our architecture is divided into 2 sub-modules. In the first part, visual and textual features are computed through a separate multi-head layer. This phase is effective enough to mingle the visual and textual features together.

In particular, with each head h in a multi-head attention block, the attentive process is similar to self-attention(SA) calculated by Equation 1 and Equation 2.

$$f_R^h = SA(f_R, f_R, f_R) \tag{1}$$

$$SA(q,K,V) = softmax(\frac{qW^Q(KW^K)^T}{\sqrt{d}})VW^v \tag{2}$$

With a stack of many self-attention layers, the visual and textual features are activated by observing the in-context components and created more powerful representation in Equation 3 and 4.

$$f_I' = Multihead(f_I) \tag{3}$$

$$f_Q' = Multihead(f_Q) \tag{4}$$

Then, the image and language signals are combined with their original features and the other's original features by a vector operation. The final signal is normalized to produce intensified visual and textual features. Although this change seems a little small, its effect is highly valuable in performance. The whole process is calculated via Equation 5 and Equation 6:

$$f_{IQ} = Norm(G(f_I', f_I, f_Q)) \tag{5}$$

$$f_{QI} = Norm(G(f_Q', f_I, f_Q)) \tag{6}$$

Where $G(.,.,.)$ is the feature combination operation. In our model, $G()$ function is chosen to achieve the best performance in Vietnamese VQA. Through our proposed combination, the features of each component are bi-directionally enhanced by the original information of the image and language. Therefore, our architecture can simultaneously and mutually learn textual and visual features.

In the second part, enhanced visual and textual features are handled by a fully connected and normalization layer similar to Transformer's Attention architecture. The final image and text signal is calculated by the Equation 7- 10.

$$f_{IQ}' = W^T f_{IQ} + b \tag{7}$$

$$f_{QI}' = W^T f_{QI} + b \tag{8}$$

$$\hat{f_I} = Norm(G(f_{IQ}', f_{IQ})) \tag{9}$$

$$\hat{f_Q} = Norm(G(f_{QI}', f_{QI})) \tag{10}$$

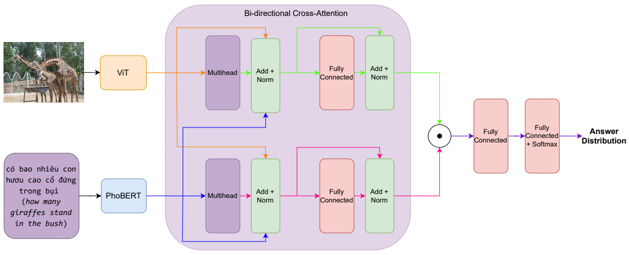

Bi-directional Cross-Attention
ViT
có bao nhiêu con hươu cao cổ đứng trong bụi (how many giraffes stand in the bush)
PhoBERT
Multihead
Add + Norm
Fully Connected
Add + Norm
Multihead
Add + Norm
Fully Connected
Add + Norm
Fully Connected
Fully Connected + Softmax
Answer Distribution

Figure 3: VQA Model System

3.3 Our model

After extracting and combining the features in the previous steps, we integrate the Bi-directional Cross-attention module into our completed system. The details of our models are presented in Figure 3. After visual and textual features are mutually augmented through the Bi-directional Cross-attention module, we only utilize simple vector operation to combine visual and textual features. Similar to most VQA approaches, we consider VQA as a classification task. So the answer distribution is calculated by the Softmax function via Equation 11.

$$y = Softmax(W_a^T.K(\hat{f_I}, \hat{f_Q}) + b_a) \tag{11}$$

Where $K(.,.)$ is the feature fusion operation. In our model, we consider $K(.,.)$ as the vector operations including addition, multiplication, and concatenation.

Corresponding to the classification approach, we use Cross-Entropy Loss as the loss function. The cross-Entropy function is calculated as follows:

$$L(y,p) = -(y\log(p) + (1-y)\log(1-p)) \tag{12}$$

$$p = Probability(y = 1) \tag{13}$$

4 Experiments

4.1 Dataset

To solve the problem of Visual Question Answering for Vietnamese, our model is evaluated on the ViVQA dataset. The ViVQA dataset is extracted and translated from the MS COCO dataset. Particularly, to avoid ambiguity when translating directly from other languages, the question-answer pairs are selected and adjusted based on the unique characteristics of the Vietnamese language. Therefore, the dataset is more accurate and realistic than using the translated data sets from English. Thus, this dataset is considered a benchmark in the VQA task for Vietnamese. The detail of the dataset is shown in Table 1.

	Train	Test
No. Samples	11999	3001
Longest Question Length	26	24
Longest Answer Length	4.0	4.0
Average Question Length	9.49	9.59
Average Answer Length	1.78	1.78

Table 1: Detail of ViVQA dataset

4.2 Experimental Setting

In our model, we leverage the power of pre-trained models in extracting visual and textual features. Besides, we also propose a Bi-directional Cross-Attention architecture to effectively improve signal learning from images and languages. Details of the configurations of each module are presented in Table 2 to facilitate the future reproduction process. Due to limited infrastructures in the experimental environment, we use some default parameters for configuration. However, through experimental results, these parameters are effective enough to implement a compact and high-performance VQA system.

4.3 Evaluation

Similar to previous VQA works, we use an accuracy score for model evaluation. With the ViVQA dataset, each question only corresponds to one answer, so the accuracy score is calculated based on the number of questions correctly answered on the entire number of questions. Let N be the number of samples, the accuracy score is calculated via Equation 14.

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{y}_i == y_i\} \qquad (14)$$

4.4 Result

Because the model built for the Visual Question Answering problem is specifically for Vietnamese, there are no published works to solve this problem. Therefore, we evaluate and compare our model with the 3 best models proposed by the author in the ViVQA dataset paper. All three models use Resnet (He et al., 2016) for visual representations and PhoW2Vec for textual features. First, we compare the model using Long Short Memory(LSTM) (Antol et al., 2015) for associative attribute learning between images and questions. Besides, we also compare the model with the model using Bidirectional Long Short Memory(Bi-LSTM) (Schuster and Paliwal, 1997) architecture for learning textual and visual features. Finally, the results of our model are compared with the model using the Hierarchical Co-Attention (Lu et al., 2016) . Details of the result are shown in Table 4.4. Obviously, our model obtains remarkable results against the existing approaches. First, our model outperforms others by using pre-trained transformers for both visual and textual representation. Second, the performance of bi-directional cross-attention architecture in concurrent learning between images and questions is more effective than LSTM, Bi-LSTM, or Hierarchical Co-Attention. This demonstrates the strength of the bi-directional cross-attention compared to previous models.

Component	Value
Vision Transformer	google/vit-base-patch16-224-in21k
PhoBERT	vinai/phobert-base
No. Cross-Attention Layer	1
No. Head	12
Dropout	0.1
Fully-connected layers	768 - 512 - 353
Optimizer	AdamW(lr = 3e-5, eps = 1e-8)

Table 2: Detail of Component Setting

5 Discussion

To demonstrate the effectiveness of each component in our proposed model, we have analyzed them through some experiments.

First, we evaluate the strength of the Bi-directional Cross Attention module compared to Transformer Attention. Besides, we also clarify the

	Accuracy
Hierarchical Co-Attention + PhoW2Vec	34.96
LTSM + PhoW2Vec	33.85
Bi-LTSM + PhoW2Vec	33.97
Our model	**51.3**

Table 3: Detail our model results against other competitive baselines

contribution of our proposed attention and its modified variants. Details of the results are shown in Table 4. Obviously, our module outperforms the original Transformer Attention through its simultaneous and mutual learning between visual and textual features. Among the different configurations, our model achieves the most promising performs when $G()$ function in Equation 5, 6, 9, and 10 is equal to the addition operation. In most variants, our proposed attention also outperforms the system without Bi-directional Cross-Attention.

	Accuracy
No Bi-directional Cross-Attention	38.5
Transformer Attention	48.6
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Add** **+ Equation 9, 10: G() = Add**	**51.3**
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Multiply** **+ Equation 9, 10: G() = Add**	38.7
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Add** **+ Equation 9, 10: G() = Multiply**	50.2
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Multiply** **+ Equation 9, 10: G() = Multiply**	35.8

Table 4: The effect of bi-directional cross-attention in our architecture

Besides, we also compare the effect of vector operations: multiplication, concatenation, and addition in combining image and question features (Equation11). The choice of multi-model fusion function makes an important contribution to improving system efficiency. Despite the simplicity of vector operation, its effect is impressive in the whole system. The results of this comparison are shown in Table 5. Obviously, the multiplication operation gives the most promising results than other operations in our model.

Operation	Accuracy
Equation 11: K() = Add	46.9
Equation 11: K() = Concatenate	37.8
Equation 11: K() = Multiply	**51.3**

Table 5: The effect of fusion operation in our architecture

Finally, we conduct experiments in selecting the number of Bi-directional Cross-Attention blocks. In our system, this module plays a critical role in the model's performance. The effect of the Bi-directional Cross-Attention block is visualized in Figure 4. The change in the model's accuracy reflects the dependency of our system on the Bi-directional Cross-Attention Block. In this experiment, our model performed best with one Bi-directional Cross-Attention block. Because of the small number of samples, It is less effective to optimize the stacked structure of attention blocks.

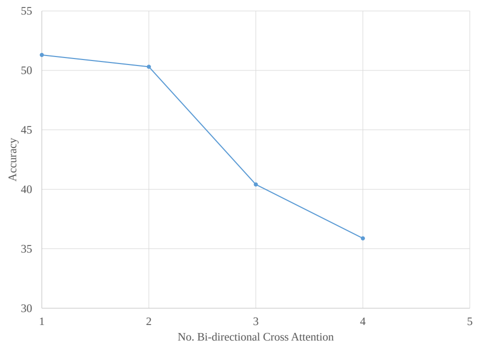


Figure 4: The effect of the number of Bi-directional Cross-Attention in our model

6 Conclusion

In this work, we propose a Bi-directional Cross-Attention Network that was fine-tuned specifically for the Vietnamese language. Our model leverages the power of pre-trained models in both Vision and Language to optimize the model's feature extraction. In particular, we adjusted the architecture of Transformer's Attention in the fusion of visual and textual features. This attention allows the image and question features to be learned simultaneously in the Transformer block. Through experiments and ablation studies, our model is proved to be more effective than other competitive baselines in Visual Question Answering for the Vietnamese language. Besides the contributions in Visual Question Answering, we also consider the extension in scientific documents where the images and figures are surrounded by many contexts in the near future works.

Acknowledgment

This work is funded by the University of Science, VNU-HCM under grant number CNTT 2022-05.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin-Shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In Helen Meng, Bo Xu 0011, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4168–4172. ISCA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Tung Le, Nguyen Tien Huy, and Nguyen Le Minh. 2020. Integrating transformer into global and residual image feature extractor in visual question answering for blind people. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 31–36.

Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. 2021a. Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing*, 465:451–464.

Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. 2021b. Vision and text transformer for predicting answerability on visual question answering. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 934–938.

Tung Le., Khoa Pho., Thong Bui., Huy Tien Nguyen., and Minh Le Nguyen. 2022. Object-less vision-language model on visual question classification for blind people. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART,*, pages 180–187. INSTICC, SciTePress.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. Vivqa: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 546–554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

The Development and Assessment of Pattern Matching Algorithms Used by ZEE: A Filipino Sign Language (FSL) Dictionary and English-Learning App

Maria Dominique Daculan
University of San Carlos
16100499@usc.edu.ph

Asherrie Jaye O. Tan
University of San Carlos
18102538@usc.edu.ph

Angie M. Ceniza-Canillo
University of San Carlos
amceniza@usc.edu.ph

Abstract

"The Filipino Sign Language Act" of RA 11106 states the declaration of Filipino Sign Language (FSL) as the National Sign Language of Filipino deaf citizens. However, the lack of awareness prevails as with the cyberbullying case of Mininio Buhat. This paper takes an initial step into raising awareness with Zee, with the analysis of three different pattern-searching algorithms - Knuth-Morris Pratt, Boyer-Moore, and Robin Karp to determine which suits the application best.By implementing three different algorithms, an efficient algorithm that is consistent in all test cases was determined to be used in pattern matching. The search functionality of the application was tested with one algorithm at a time where users input an English word and the application searches the database returning the corresponding FSL gesture of that particular English word in gif format. From there, the execution time of each algorithm was recorded and the overall criteria for the analysis of the algorithms: time complexity, space complexity and overall performance was compared among the three algorithms. It was determined that the Boyer Moore algorithm gave consistent results and was generally fast and efficient in any test case, be it patterns of shorter length and patterns of longer length.

1 Introduction

According to a survey, the term disability refers to "any restriction or lack of ability (resulting from impairment) to perform an activity in the manner or within the range considered normal for a human being" (NSO, 2014). Focusing on those belonging to the deaf community, Center for Hearing and Communication statistics show that around 90% of deaf children are born to hearing parents.

In the Republic Act 7277 [1], otherwise known as Magna Carta for Disabled Persons, it is stated that people who are deaf, mute, or hearing-imapired belong to 1.23% of the entire Philippine population. Furthermore, the 2009 projected population of the deaf consists of 241,624 totally deaf citizens as well as 275, 912 partially deaf. In the aforementioned Act, Persons with Disabilities (PWD's) are given special privileges and self-development programs for them to be a part of the mainstream of society. Despite having these laws made to alleviate the burden PWD's carry, it is eminent in society that people lack awareness towards PWDs and would often regard them as "weird", "abnormal" and in most cases "special" in a negative nuance. One of the major factors contributing to these preconceived ideas is the lack of communication and understanding between the hearing and the non-hearing community.

In addition, a 2014 Philippine News article presented Mininio Buhat - a Filipina born profoundly deaf - who became a victim of cyberbullying shortly after posting on the social platform Facebook with faulty written English. It was later revealed by experts that deaf people are taught proper English grammar in the latter part of their college education, as well as the fact that her sentence construction was based upon their sign language. This further proves the lack of awareness and knowledge the hearing majority has regarding the deaf and how their language works for them to communicate. According to the list by the National Council on Disability Affairs (NCDA), there exists a number of organizations and schools offering Special Education (SpEd) that share the advocacy of inclusivity - giving education to the SpEd individuals as well as providing FSL training to both the deaf and the hearing. Mandaue City Central Special Education School (MCCSpEd), FSL National Network, Philippine Accessible Disability Services Inc.(PADS) are just a few of those that are actively expanding their advocacies to society. However, people don't seem to be aware of such efforts and still disregard the importance of their Advocacy.

With that in mind, this study proposes Zee. It is an application that promotes interactive learning of FSL for the hearing, as well as English grammar to the deaf. Through Zee, the advocacy of inclusivity is shared between the application and its users, in hopes of closing the communication gap between the deaf and the hearing community as well as raising awareness regarding the importance of FSL in our society. In lieu of this, an assessment of the aforementioned pattern-searching algorithms is proposed to determine which suits best for the type of dataset Zee has.

This research aims to assess the efficiency of the aforementioned pattern-searching algorithms to be used in the development of Zee: A Filipino Sign Language (FSL) Dictionary and English Learning App.

The objectives of the study are as follows:
1. Gather information on the current situation of the deaf community, their education, and their relationship with the mainstream society
as well as evaluating the current applications available for learning FSL.
2. Design and develop an FSL and english learning application using pattern matching algorithms.
3. Assess the pattern matching algorithms used in developing the application.
4. Test and evaluate the performance of the application.

The following will benefit from the results of this study:
Deaf Community. This study would be of help to the said community as it will teach them to construct basic English sentences in the proper grammar as early as their high school years, as proper grammar is said to be taught only in the later part of their college education. This study would also raise awareness of the community's advocacy for inclusivity.

Families with deaf-mute relatives and friends.Similarly, families will benefit from this study as they will now be able to learn the language used by their deaf-mute relatives and acquaintances and would be able to start communicating with them and gain a bit more knowledge and understanding towards the deaf-mute community.

Teachers of SpEd Schools. These teachers have been the pioneers of the advocacy of inclusivity by giving them the education they need and deserve, as they prepare the students for living in society. Through this study, it would support and act as gratitude for the efforts of these teachers by providing the students a platform to learn, especially in this time of pandemic and online learning.

Schools offering SpEd.Similar to the teachers, it is through the schools that deaf students are given a comfortable learning environment wherein SpEd and non-SpEd students can interact with each other and can lessen the 3 language barrier that exists between them. This study would benefit these institutions as it reflects their advocacy as it allows the deaf and the hearing community an opportunity to minimize the language barrier that exists.

Hearing Community. As sign language resources are a challenge to find, this study would provide the hearing community an opportunity to learn sign language - the official language of the deaf-mute community. Through this, they would be able to understand those that are deaf-mute and would be able to communicate with them even through basic conversations.

Researchers. The researchers will benefit from this study because this allows them to gain more knowledge and experience in creating an application, as well as knowing the best suited algorithm for this type of application.

Future Researchers. The study will benefit future researchers on their research on similar topics.

This study is for the assessment of different pattern-matching algorithms (Boyer-Moore, Rabin-Karp, and Knuth-Morris Pratt algorithm) for the development of a web, mobile-responsive application for learning FSL and English. This teaches the deaf community in their high school years to construct basic English sentences according to the correct grammar rules. In addition, this application brings forth awareness of the deaf community and serves as an opportunity for us - the hearing community - to learn the different signs and to be able to conjure a conversation with a deaf or hard of hearing individual. Sign Languages consist of manual (hand movements) and non manual signals (head nod, eye blink, etc.). In this study, only the manual signals are considered, with slight inclusion of non manual signals concerning signs about questions.

The different signs will be in a form of Graphics Interchange Format (GIF). Each sign will be demonstrated by Elizabeth Ann Daculan, a high school deaf student. The demonstration clips will then be sent to the contacted advisers of Mandaue City Central Special Education School to validate the GIF clips. The 4 proposed study is primarily a web application, accessible through either Google Chrome, Microsoft Edge, and Mozilla Firefox. Furthermore, Zee will also be mobile responsive; hence, can be accessed through a mobile browser.

The Dictionary module consists of the different FSL signs in different categories, along with their corresponding meaning and guidelines on how to do the sign. In addition, users would also be able to search certain words (if available). The content is taken from a scanned reference book owned by the researchers in 2019; furthermore, the ones included in this study will be limited to the Preparatory level English signs which consists of the basic signs (ordinal and cardinal numbers, color, family, common phrases and vocabulary), the Filipino alphabet.

The application tracks the progress of its users - both the deaf and hearing - through different categories wherein each category has several levels of difficulty and with assessment tests after every level in the grammar module. On one hand, the categories for the hearing users run from the Filipino alphabet to basic vocabulary pertaining to color, numbers, family and relatives, days of the week and months, common phrases, and ultimately the Philippine National Anthem. Assessments are done through matching-type and/or fill in the blank type of questions.

On the other hand, categories for the deaf users cover the four basic sentences - Declarative, Interrogative, Imperative, and Exclamatory sentences. All sentences will be in the active voice for simplicity and ease of understanding. When relating the signs to the sentence, each pair of signs and its corresponding word in the sentence will be color-coded for identification. Similar to how the hearing users are assessed, matching type and fill in the blanks are the questions that will be used for assessment.

A review quiz will be given to the users after every level. Should the user not be able to get everything correct, he/she is not allowed to move to the next level and must restart the level. A successful attempt on the review quiz rewards the user with a star, and the total number of stars to collect per category is equal to the number of levels it has. The assessment tests will be given after every category - containing jumbled questions from the different levels the category contains. Similar to the review quiz, a user is not able to continue to the next category until he/she has passed the assessment test.

2 Related Works

2.1 The Deaf Community

In a research by Calimpusan and Silva-dela Cruz (2018), those belonging to the Deaf community are defined as people with a “hidden disability”, as they lack physical marks and attendants (wheelchair, walking cane, black eyeglasses, service dogs, etc.) as opposed to other disabilities. The only recognition they get is when they begin communicating using sign language.

Cabreros (2020) studied on the English proficiency level of 23 college Special Education (SPED) students taking the Associate in Computer Technology program under the Bachelor of Science (BS) in the Information Technology department. Results showed that a large percentage of the respondents scored below the mean score, given an interpretation of Needs Improvement. The average result determined that most of the respondents portrayed below average skills on their English proficiency in terms of their vocabulary. Respondents gained a significant improvement on their English grammar, but portrayed low comprehension skills.

2.2 Filipino Sign Language (FSL)

Several studies, including one by Rinaldi, Caselli, Lucioli, Lamano, and Volterra (2018), have emphasized the difference between Signed Languages (SLs) and spoken languages. One of which being the difference between their basic linguistic units. Spoken languages sequentially produce phonemes - the smallest unit of sound in a word that makes a difference in its pronunciation, as well as its meaning (LiteraryDevices) - and morphemes - the smallest syntactical 7 and meaningful linguistic unit that contains a word (LiteraryDevices) - whereas signed languages produce them simultaneously. Signed languages are also divided into two parameters, the manual parameters (movement, hand location, orientation) as well as the non-manual parameters (facial expression, mouthing, body movements) which gives a different meaning to every sign used when communicating.

The aforementioned study also did an experiment with three different deaf age groups (younger children, older children, adults) and tasked them to reproduce the sign language shown on the computer. Results showed that younger children tend to omit a larger number of signs then the other groups, while the remaining groups did not show any significant differences. Younger children presented a pattern of reproducing the signs different from the older children and adult groups.

Signed by President Duterte on the 30th of October 2020, the Republic Act 11106 - otherwise known as The Filipino Sign Language Act - officially recognizes FSL as the National Sign Language and be officially used in all types of transactions with the Deaf community. Section 11 of said Act states the promotion of FSL by the appropriate agencies towards propagating the hearing people's competency in learning FSL by offering it as an elective in the mainstream curriculum, particularly of the State Universities and Colleges (SUCs) or schools that are government-funded. However, a section that talks about family members learning FSL in conjunction with their deaf child/children's learning is not present.

Uploaded on Google Playstore in February of 2018, Filipino Sign Language is a simple informative program that teaches sign-language gestures or "fingerspell" of the basics (letters, numbers, special characters) and offers an FSL translation of the Philippine National Anthem. Considered as the first of its kind due to similar apps only up to its first release and has never been updated.

Another application named FSL Buddy was uploaded last July 2018. The goal of the app - as well as other apps similar to them - have a goal of teaching people the national sign language. The FSL vocabulary featured in this app is 8 part of a course taught at De La Salle – College of Saint Benilde. This app was only up to its first release. Recent applications lack the feature of teaching users how to use the signs in a sentence.

2.3 Sign Language Translation

In this study by Orbay and Akarun (year), a pre-processing method called tokenization was used as well as utilizing token learning from sign language videos wherein supervised data is at hand. As annotated data is costly and scarce altogether, Transfer-Learning from semi-supervised tokenization approach is utilized. Three experimental setups were determined; right hand keypoints, both hand keypoints, as well as full body pose while ignoring the non-manual signs (face keypoints). The said study utilizes RWTHPHOENIX-Weather 2014T which is a Continuous Sign Language Benchmark dataset. This study only focuses on the manual portion of the signs and did not put into consideration the non-manual portion as well as the context of the sign that would help produce better results. This study does not in any way discuss the linguistic properties of sign language. Another study by Camgoz, Hadfield, et. al., utilizes the Neural Machine Translation (NMT) framework inorder to determine the spatial representations, its language model, as well as the mapping in between the sign and spoken language.

In this study by Orbay and Akarun (year), a pre-processing method called tokenization was used as well as utilizing token learning from sign language videos wherein supervised data is at hand. As annotated data is costly and scarce altogether, Transfer-Learning from semi-supervised tokenization approach is utilized. Three experimental setups were determined; right hand keypoints, both hand keypoints, as well as full body pose while ignoring the non-manual signs (face keypoints). The said study utilizes RWTHPHOENIX-Weather 2014T which is a Continuous Sign Language Benchmark dataset. This study only focuses on the manual portion of the signs and did not put into consideration the non-manual portion as well as the context of the sign that would help produce better results. This study does not in any way discuss the linguistic properties of sign language. Another study by Camgoz, Hadfield, et. al., utilizes the Neural Machine Translation (NMT) framework inorder to determine the spatial representations, its language model, as well as the mapping in between the sign and spoken language.

Lugman and Mahmoud (2018), developed a semantic rule-based machine translation system from Arabic into Arabic Sign Language (ArSL). This system involves three main translation stages (morphological analysis, syntactic analysis, and ArSL generation. The user inputs an Arabic sentence. The system then outputs an ArSL sentence represented by the gloss notation which is displayed as a sign sequence of GIF images. The sentence is first morphologically analyzed using the MADAMIRA toolkit to extract the structure of each of its words. This output is then syntactically analyzed using a dependency parser to obtain the relation between the words in the sentence. An Arabic parse tree is generated and then transformed into its equivalent tree in ArSL by applying some transformation rules. This phase also involves lexically translating the Arabic phrases and words into their equivalent signs in ArSL. In the event of an out-of vocabulary (OOV) problem, the synonym of the word is used. A statistical language model in the synonym selection is used to ensure that the synonym preserves the meaning of the sentence.

3 Methodology

This study consists of two user categories - deaf and hearing. In terms of the deaf respondents, the researchers will be collaborating with Ms. Dioscora Sollano, an adviser of the Hearing Impaired (HI) department of Mandaue City Central Special Education School. Considering the current online setting, the exact number of respondents for the deaf category cannot be determined, as there exists the challenge of contacting the parents of the students as well as the availability of internet connection in each household.

For the hearing respondents, the 155 BS Computer Science (BSCS) students from the Department of Computer, Information Sciences and Mathematics of the University of San Carlos - Talamban Campus is our chosen population.

$$n = \frac{N}{1 + N(e)^2}$$

Figure 1. Yamane Formula

Figure 1 shows how the sample size from the 155 BSCS students are calculated. n represents the sample size, N represents the total population - in this case, 155 -, while e is the margin of error which is set to 0.05 based on the research conditions. These values would then generate a sample size of 111 BSCS students.

For the initial information gathering, two sets of researcher-made questionnaires; one for the deaf respondents and the other for the hearing respondents. The questions are structured in a way to acquire qualitative data on their opinion on the awareness of the deaf community in society. Aforementioned questions are inspired from the personal experiences of one of the researchers, news articles, past interviews, and conversations with people directly related to the deaf community.

In the Knuth-Morris Pratt algorithm, let w be the substring and s be the given string and is the basis of comparison. A one-dimensional array (variable arr) is utilized and initialized according to the number of characters that can be skipped after a mismatch is met in every iteration. For the initialization of arr, a failure function f(i) is used. Such function is basing its procedure on the fact that all previous characters are correctly matched once a mismatch occurs. Hence, when a w prefix occurs in the set of matched characters, it is also a w suffix. In the function proper, a string with a length of 1 is given the value of 0 to arr.

For the Robin-Karp algorithm; let pattern be the string to be searched, m be the length of the pattern, text as which the pattern will be searched from, as well as N to be the number of characters. To start, the hash value of pattern is calculated and utilized later for comparison in pattern_hash. M characters from text are then calculated of their own hash value in text_hash. Both values of pattern_hash as well as text_hash are then compared; if they are equal, a brute force comparison is then executed wherein individual characters are compared from pattern and text, otherwise the next M characters of text are then calculated for their hash value and comparison begins. The process ends once the end of text is reached or the brute force comparison results in a TRUE.

In the case of the Boyer-Moore algorithm; let text be the origin string, pattern be the string to be searched, and N being the length of the pattern . Beginning 16 from the rightmost character of pattern (place variable i), compare that character with its equivalent placement in text (place variable j). If those initial characters mismatch, and if the character from text is not detected anywhere in pattern, pattern is then shifted N characters to the right. But if that character is detected, the pattern is shifted until the occurrence of that character in the pattern is aligned with the character in text. This process continues until all characters have a match or if it reaches the end of text.

Other sources such as websites, videos are used for data gathering. For the in-application assessment, a pre-test and post-test is done after each category in the application. The same questions are used in both tests to properly compare the change of scores. Evaluation is then done through the use of a T-Test.

By having respondents in their high school years involved, the researchers will apply for a research ethics review. Any information gathered from the respondents will be treated confidential between the researchers and the respondents themselves.

Both qualitative and quantitative approaches are used in this study. A qualitative approach is used in the initial data collection of the study to demonstrate the current awareness of the respondents regarding the deaf community in society. Letters were sent to the respective school heads to obtain permission on having a few of their students as respondents to the study. Once approval is received, contact information of the respondents are then acquired, and questionnaires will be transformed into Google Forms wherein the links will be sent to each of the respondents.

For the quantitative approach, a pre-test is administered before every category of the application. Individual scores are then recorded to be used in analysis. After which, respondents will then undergo a series of lessons in the form of levels and taking quizzes in between. Their progress per test is also recorded. After each category, a post-test is administered having the same questions as the pre-test. Results of the post-test are also recorded and will be analyzed together with the pre-test results.

A t-test is a statistical test used to compare the means of two groups, in this case, between the deaf and hearing respondents. It is often used to determine whether or not a process or treatment is effective on the population of interest. (Bevans, 2021)

Paired t-test is specifically used in this study to determine the mean difference of the English proficiency test scores taken by the respondents before and after learning each lesson category in the application. The results will then be analyzed and presented through graphs. SPSS software will be used for this purpose. ("LibGuides: SPSS Tutorials: Paired Samples t Test", 2021)

$$s_{\overline{x}} = \frac{s_{\text{diff}}}{\sqrt{n}}$$

Figure 2. Paired t-test Formula

Figure 2 shows the formula for the Paired t-test. Sdiff represents the sample standard deviation of the differences, sx¯ represents the estimated standard error of the mean (s/sqrt(n)) and n represents the sample size (i.e., number of observations).

Three different pattern-searching algorithms are utilized in this study, namely the Knuth-Morris Pratt, Boyer-Moore, as well as the Robin-Karp algorithm. Each of these algorithms will be implemented in the Dictionary module and analyzed based on a number of factors. To analyze each of the three algorithms, a variable O is defined to represent the running time and another variable n is defined to represent the space complexity. Two other variables Sn, representing the number of characters skipped from the input string in the event of a mismatch and Cn, are defined to represent the number of comparisons done overall.

In evaluating algorithm performance, the following factors need to be considered: correctness, finiteness, and efficiency. An algorithm is said to be correct if it produces the desired output for any particular set of inputs. It is vital that an algorithm should terminate after a finite number of steps. Otherwise, should loops be used, an infinite loop will occur and the program cannot progress until the error is resolved by the programmer. Lastly, an algorithm should be efficient meaning it should take up as less time and use less space, memory and resources as possible. ("Time and Space Complexity Analysis of Algorithm", 2019)

To further compare the three algorithms, certain formulae will be used to determine their respective efficiency and correctness. In measuring algorithm efficiency, the Big-O notation is used to determine its running time. The general formula for getting the running time is O(n)=cn, where c is some variable that represents the number of statements executed by the program. Using this formula, time complexity is calculated depending on the number of operations and the size of the input. Below shows the running time for different types of operations performed:

O(1) - when the algorithm performs a constant number of operations regardless of the size of the input.

O(log n) - when the algorithm takes as many steps as it can by performing repeated operations, for example division.

O(n) - for every loop executed
("Running Time and Big-O - Learneroo", n.d.)

An algorithm is considered to be correct whenever the initial state fulfills the precondition and the program terminates. Thus the formula for total correctness is:

Total Correctness = Partial correctness + Termination (Tran, 2017)

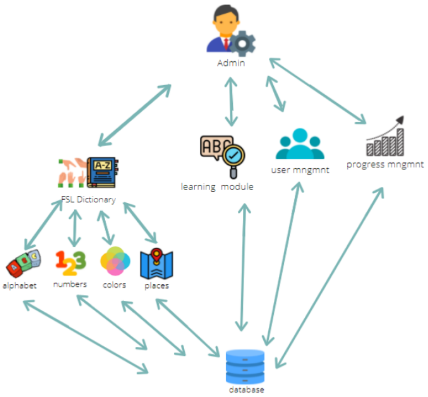


Figure 3 Administrator Conceptual Framework

Zee is an English language learning web and mobile-responsive application that aims to improve the communication between the deaf community and the hearing community. The learning application consists of several modules: an FSL Dictionary and a grammar module. The grammar module is subdivided into two - an FSL to sentence and a sentence to FSL submodule. The users of this application are the administrators, the deaf-mute community, and the hearing community. The administrators handle the database and updates whenever needed as well as accessing other modules. Both communities can access all modules but are unable to access and update the database itself.

As shown in Figure 3, the administrator can access all the modules, input and edit data into the database. The dictionary module is used in the learning module as a resource for lectures and tests, as well as allowing users to view the different signs. A Graphic Interchange Format (GIF) is used to creatively display how each sign is done as well as a textual description of how to act the sign. Each category of signs is assigned a color for ease of identification, which will be useful on the second module.

On the other hand, the learning module - separated into two submodules - aids deaf-mute users in constructing sentences in the basic types (declarative, interrogative, imperative, exclamatory) in the English grammar system through means of a game that assesses their knowledge in between. The FSL equivalent word in the sentence will be tagged according to the sign's category in the dictionary. This same means is applied to the hearing users to understand how the different signs are done and used in a sentence.

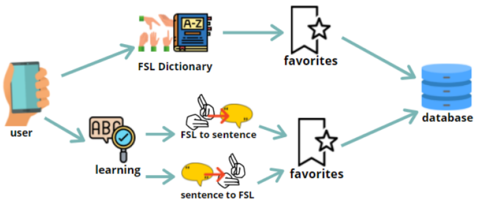


Figure 4 User Conceptual Framework

Users can access all modules in the application as shown above in Figure 4. When accessing each module and submodule, users can opt to add certain items - in the dictionary and learning modules - into Favorites to save for future access.

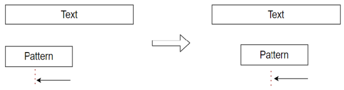


Figure 5 Boyer Moore algorithm

The Boyer-Moore Algorithm, specifically the Bad-Character Rule shown above, works this way. A given pattern is checked with the beginning of a certain string/text. Checking is done from right to left to see if any mismatched 22 characters exist between the two strings. If a mismatched character is found, the pattern moves one character past an occurrence of the mismatched character to the left. If no occurrence of the mismatched character is found to the left, the pattern moves one character to the right of the mismatched character. From there, checking continues until the end of the string is reached. (Langmead, 2015)

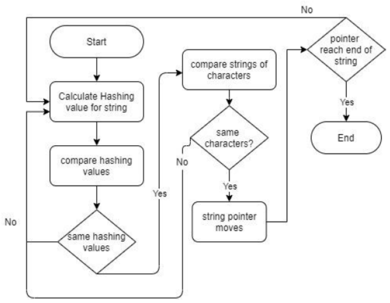


Figure 6 Rabin Karp Algorithm

The Rabin-Karp algorithm in the figure above makes use of hash functions and associates each character with a value from these functions. Pattern matching is done by adding the hash values of each character from a given pattern and doing the same for the entire string to be checked. If the sum of the hash values of both patterns are not equal, a mismatch occurs and the pattern moves one space to the right. The process repeats until the end of the given string is reached. (Bari, 2018)

The figure below shows how the Knuth-Morris Pratt algorithm works. It starts by initially setting traversal variables to the beginning of both the pattern and given string. When a mismatch is met, the pattern traversal variable goes back to the beginning of the pattern to restart comparison. The algorithm 23 continues until the end of the given string is reached, whether or not a match is found.

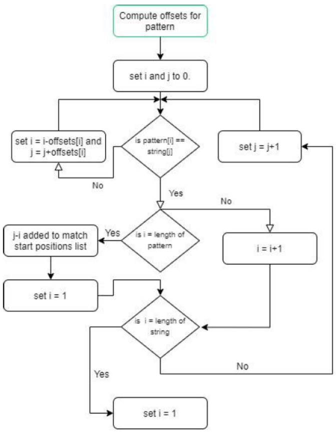

Compute offsets for pattern
set i and j to 0.
set i = i-offsets[i] and j = j+offsets[i]
is pattern[i] == string[j]
set j = j+1
No
Yes
No
j-i added to match start positions list
Yes
is i = length of pattern
i = i+1
set i = 1
is i = length of string
Yes
No
set i = 1

Figure 7 Knuth Morris Pratt Algorithm

According to Figure 8, the lexical table entitled ‘entries’ is taken from the “MySQL English Dictionary” which was parsed from the OPTED database and does not have any relation with the other tables. Every FSL sign corresponds to only one word in the entries table. In every favorite, there could be one or more FSL signs, and can be marked by one or more users. In terms of the progress, each entry of progress is to only one user, while every user can have one or more progress entries.

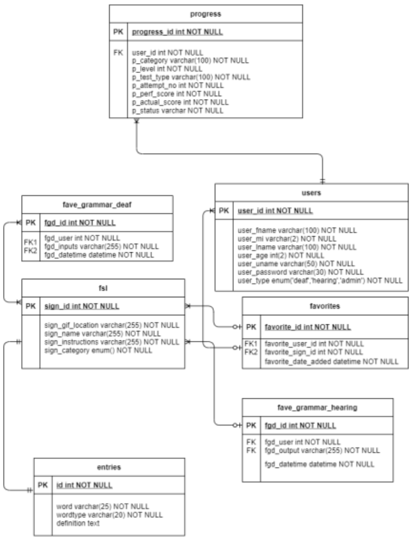

progress
PK
progress_id int NOT NULL
FK
user_id int NOT NULL
p_category varchar(100) NOT NULL
p_level int NOT NULL
p_test_type varchar(100) NOT NULL
p_attempt_no int NOT NULL
p_perf_score int NOT NULL
p_actual_score int NOT NULL
p_status varchar NOT NULL
fave_grammar_deaf
PK
fgd_id int NOT NULL
FK1
FK2
fgd_user int NOT NULL
fgd_inputs varchar(255) NOT NULL
fgd_datetime datetime NOT NULL
users
PK
user_id int NOT NULL
user_fname varchar(100) NOT NULL
user_mi varchar(2) NOT NULL
user_lname varchar(100) NOT NULL
user_age int(2) NOT NULL
user_uname varchar(50) NOT NULL
user_password varchar(30) NOT NULL
user_type enum('deaf','hearing','admin') NOT NULL
fsl
PK
sign_id int NOT NULL
sign_gif_location varchar(255) NOT NULL
sign_name varchar(255) NOT NULL
sign_instructions varchar(255) NOT NULL
sign_category enum() NOT NULL
favorites
PK
favorite_id int NOT NULL
FK1
FK2
favorite_user_id int NOT NULL
favorite_sign_id int NOT NULL
favorite_date_added datetime NOT NULL
fave_grammar_hearing
PK
fgd_id int NOT NULL
FK
FK
fgd_user int NOT NULL
fgd_output varchar(255) NOT NULL
fgd_datetime datetime NOT NULL
entries
PK
id int NOT NULL
word varchar(25) NOT NULL
wordtype varchar(20) NOT NULL
definition text

Figure 8. Entity Relationship Diagram

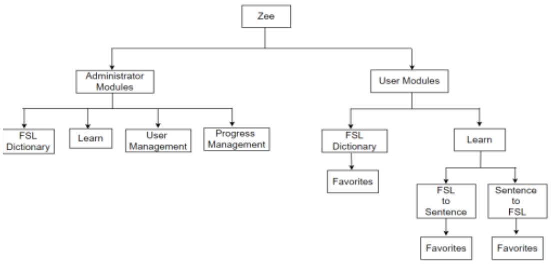


Figure 10. Top-Down Approach

This project utilizes the top-down approach for the development of the application. The approach allows a more defined state of development in every step as undefined submodules are defined after each step.

4 Results and Findings

The three pattern matching algorithms were implemented in the Zee FSL to English dictionary. Each algorithm was implemented one at a time as the search feature was tested. Users are to input a certain English word or pattern where the algorithm searches for a match in the database consisting of 201 entries and returns the corresponding Filipino sign language gesture with the corresponding gif format.

Each algorithm was analyzed based on the following metrics: Time Complexity, Space Complexity and Performance (Execution time).

pattern	Knuth Morris	Rabinn Karp	Boyer Moore
teen	6527	5017	5008
ther	6601	5009	5012
ch	6584	5020	5013
o	6594	5076	5013
day	6518	5017	5010

Table 1 Execution of all three Algorithms in searching small partial patterns

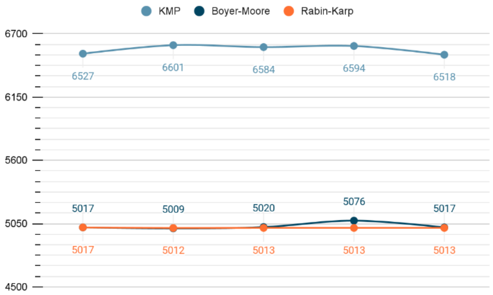


Figure 7 Screenshot to compare the execution time of the three Algorithms with searching small partial patterns

pattern	Knuth Morris	Rabinn Karp	Boyer Moore
small	6655	5006	5012
classmate	6778	5012	5005
read	6570	5007	5005
different	6717	5011	5015
why	6752	5014	5007

Table 2 Execution of all three Algorithms in searching full patterns

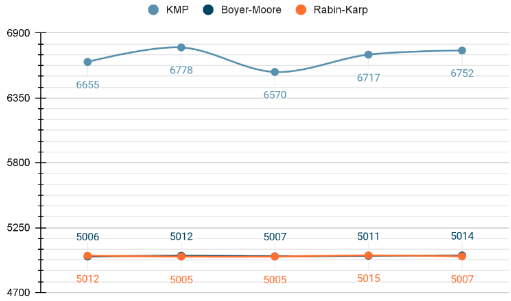


Figure 7 Screenshot to compare the execution time of the three Algorithms with searching full patterns

pattern	Knuth Morris	Rabinn Karp	Boyer Moore
invalid	6734	5005	5013
mountain	6654	5009	5009
xxx	6654	5005	5014
presents	6614	5006	5008
up	6619	5012	5016

Table 3 Execution of all three Algorithms in searching invalid patterns

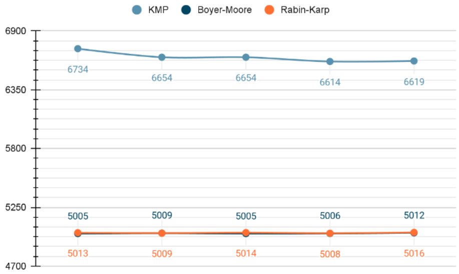


Figure 8 Screenshot to compare the execution time of the three Algorithms with searching invalid patterns

5 Conclusion and Recommendations

After running some tests, it was determined that the Knuth Morris Pratt Algorithm executes the slowest. This is considering that the words in the database are mostly of shorter length, with an average of 5 characters. Since the Knuth Morris Pratt Algorithm is designed to check for repeating characters known as prefixes and suffixes, then it behaves differently when all the characters in the string are unique..

Overall, the Boyer Moore Algorithm seems to be the most efficient. This is because the Boyer Moore algorithm is consistent in skipping a couple of characters when a mismatch is encountered.

However, when both patterns have similar subpatterns known as prefixes and suffixes, the Knuth Morris Pratt Algorithm works faster as it immediately goes to a common subpattern and begins checking from there until the last index

In terms of space complexity, the Rabin Karp Algorithm utilizes the least amount of space compared to the two algorithms. However, this low cost space is at the expense of a high time complexity as it only shifts one character to the right at a time.

While the algorithms analyzed in the paper may have an old approach, it is encouraged that future researchers who wish to study pattern matching make their own algorithms that are hopefully faster and more efficient in checking and comparing the existence of patterns than these existing algorithms. It will be even better if the new future algorithm proposed by other researchers will work smoothly as intended and hopefully easy to comprehend. New pattern matching algorithms will pave the way for faster pattern comparisons to be used by professionals not only in the field of Computer Science but also other fields that utilize pattern matching algorithms like Bioinformatics and DNA sequencing.(Fainstein, 2005)

References

Abdallah, E., & Fayyoumi, E. (2016). Assistive Technology for Deaf People Based on Android Platform. Procedia Computer Science, 94, 295-301.doi: 10.1016/j.procs.2016.08.044.

Cabreros, M. R. Q.. English Language Proficiency Profile: A Case Study of the Communication Skills of Deaf Students in the Undergraduate Program in Quezon, Philippines. International Journal of Educational Management and Development Studies, 1(1), 20-42.

Fainstein, J., 2005. The application of pattern matching algorithms in bioinformatics. [online] (UMI 1430487). Available at: <https://www.proquest.com/openview/d206c1963532e01b391e32cdf1c8b516/1?pq-origsite=gscholar&cbl=18750&diss=y> [Accessed 20 May 2022].

Luqman, H., & Mahmoud, S. (2018). Automatic translation of Arabic text-to-Arabic sign language. Universal Access In The Information Society, 18(4), 939-951. doi: 10.1007/s10209-018-0622-8.

Rinaldi, P., Caselli, M., Lucioli, T., Lamano, L., & Volterra, V. (2019). Sign Language Skills Assessed Through a Sentence Reproduction Task. Journal of Deaf Studies and Deaf Education, 408–421. doi:10.1093/deafed/eny021.

Silva-Dela Cruz, F. W. & Calimpusan, E. C.. Status and Challenges of the Deaf in One City in the Philippines: Towards the Development of Support Systems and Socio-Economic Opportunities. Asia Pacific Journal of Multidisciplinary Research, 6(4), 61-74.

Sleit, A., AlMobaideen, W., et. al., 2007. An Efficient Pattern Matching Algorithm. Journal of Applied Sciences, 7, 2691-2695.

Suszczanska, N., Szmal, P., & Kulikow, S. (2007). Continuous Text Translation Using Text Modeling in the Thetos System, 1(8). doi: doi.org/10.5281/zenodo.1331657.

Access Innovation Media. 4 Reasons People Use Sign Language Other Than Hearing Loss. Retrieved 7 March from https://blog.ai-media.tv/blog/4-reasons-people-use-sign-language-other-th an-hearing-loss.

All About Cookies. About your web browser. Retrieved 5 May from https://www.allaboutcookies.org/browsers/.

Bevans, R. (2021). An Introduction to T-Tests | Definitions, Formula and Examples. Retrieved 5 May 2021, from https://www.scribbr.com/statistics/.

Center for Hearing and Communication. Statistics and Facts About Hearing Loss. Retrieved 7 March from https://chchearing.org/facts-about-hearing-loss/.

Codecacademy. What Is a Relational Database Management System? Retrieved 5 May from https://www.codecademy.com/articles/.

Codecacademy. What Is an IDE? Retrieved 28 April from https://www.codecademy.com/articles/what-is-an-ide.

Deafblind Support Philippines. Persons with Disabilities: Status in the Philippines. Retrieved 6 March from http://web.nlp.gov.ph/nlp/sites/default/files/.

Domantas, G. What is Web Hosting? Web Hosting Explained for Beginners. Retrieved 18 April 2021 from https://www.hostinger.com/tutorials/what-is-web-hosting/.

Educative. What is the Knuth-Morris-Pratt algorithm? Retrieved 5 June 2021 from https://www.educative.io/edpresso/what-is-the-knuth-morris-pratt-algorithm.

Farzand, A. Yamanes Formula - Sample Calculation. Retrieved 30 April 2021 from https://www.scribd.com/document/438058892/Yamanes-Formula-SampleCalculation.

FreeCodeCamp. The Rabin-Karp Algorithm Explained. Retrieved 07 June 2021 from https://www.freecodecamp.org/news/the-rabin-karp-algorithm-explained/.

Kent State University. Deaf Community Definition. (2021). Retrieved 29 April 2021 from Kent State University - Department of Modern and Classical Language Studies: https://www.kent.edu/mcls/deaf-community-definition.

Leonardo, B., Hansun, S. (2017). Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms. Retrieved 20 May from Rabin-Karp Algorithm | Download Scientific Diagram (researchgate.net).

LibGuides: SPSS Tutorials: Paired Samples t Test. (2021). Retrieved 5 May 2021, from https://libguides.library.kent.edu/SPSS/PairedSamplestTest.

Morpheme. (n.d.). In Literary Devices Definition and Examples of Literary Terms. Retrieved from https://literarydevices.net/morpheme/.

National Council on Disability Affairs. Public Elementary School. Retrieved from https://www.ncda.gov.ph/cebu/.

Opensource. What is open source? Retrieved 5 May from https://opensource.com/resources/what-open-source.

Orbay, A. & Akarun, L. (2020). Neural Sign Language Translation by Learning Tokenization. Retrieved 24 February from http://www.counseling.org.

Paired Sample T-Test - Statistics Solutions. (2021). Retrieved 5 May 2021, from https://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/.

Phoneme. (n.d.). In Literary Devices Definition and Examples of Literary Terms. Retrieved from https://literarydevices.net/?s=phoneme.

Plaxton, G. (2004). String-Matching: Rabin-Karp Algorithm. Retrieved 20 May 2021 from https://www.cs.utexas.edu/~plaxton/c/337/04s/slides/StringMatching-1.pdf.

Running Time and Big-O - Learneroo. Retrived from https://www.learneroo.com/modules/106/nodes/559.

Smith, A. (2021, April). What Are Text Editors and Why Are They Important?. Retrieved 5 May from https://www.techstuffed.com/what-are-text-editors-and-why-are-they-impor tant/.

Software Testing Fundamentals (2020). Black Box Testing. Retrieved from https://softwaretestingfundamentals.com/black-box-testing/.

Time and Space Complexity Analysis of Algorithm. (2019). Retrieved from https://afteracademy.com/blog/time-and-space-complexity-analysis-of-algo rithm.

Tran, H. (2017). How to prove correctness of algorithm. Retrieved from https://medium.com/@tranduchanh.ms/partial-correctness-of-computer-pr ogram-f541490e7a21.

Umbao, E. (2014, August 7). Mininio Buhat: A Deaf Filipino Cyber Bullied Because of Faulty Written English Grammar. Philippine News. 35 https://philnews.ph/2014/08/07/mininio-buhat-deaf-filipino-cyber-bullied-because-faulty-written-english-grammar/.

What is IDE?. (n.d.). Retrieved from https://www.codecademy.com/articles/what-is-an-ide.

Videos

Bari, A. (2018). Knuth-Morris-Pratt KMP String Matching Algorithm [Video]. Retrieved from https://www.youtube.com/watch?v=V5-7GzOfADQ.

Bari, A. (2018). Rabin-Karp String Matching Algorithm [Video]. Retrieved from https://www.youtube.com/watch?v=qQ8vS2btsxI.

DeafED Philippines (2020, August 30). FILIPINO SIGN LANGUAGE EXPLAINED [Video]. YouTube. https://youtu.be/JYR8OgSViUY.

Inquirer.Net (2019, September 23). What we ought to know about Filipino Sign Language [Video]. YouTube. https://youtu.be/3yX4GN6wG0k.

Langmead, B. (2015). Boyer-Moore basics [Video]. Retrieved from https://www.youtube.com/watch?v=4Xyhb72LCX4.

Ramesh, B. (2020). Boyer Moore Pattern Matching Algorithm [Video]. Retrieved from https://www.youtube.com/watch?v=4Oj_ESzSNCk.

Government Legislation

Magna Carta for Disabled Persons 1992. (Ph) (Ph)

S.B. 2117, Sixteenth Congress of The Republic of The Philippines, First RegularSession(2011). http://legacy.senate.gov.ph/lisdata/1868815815!.pdf.

The Filipino Sign Language Act 2018. (Ph) s. No. 1455 (Ph).

Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation

Koyel Ghosh
Central Institute of Technology,
Kokrajhar, Assam, India
ghosh.koyel8@gmail.com

Apurbalal Senapati
Central Institute of Technology,
Kokrajhar, Assam, India
a.senapati@cit.ac.in

Abstract

Warning: This paper contains examples of the language that some people may find offensive.

Transformer-based Language models have achieved state-of-the-art performance on a wide range of Natural Language Processing (NLP) tasks. This work will examine the effectiveness of transformer language models like BERT, RoBERTa, ALBERT, and DistilBERT on existing Indian hate speech datasets such as HASOC-Hindi (2019), HASOC-Marathi (2021) and Bengali Hate Speech (BenHateSpeech) over binary classification. Most deep learning methods fail to recognize a hate sentence if hate words are wrapped into sophisticated words where transformers understand the context of a hate word present in a sentence. Here, Transformer-based multilingual models such as MuRILBERT, XLM-RoBERTa, etc. are compared with monolingual models like NeuralSpace-BERTHi (Hindi), MahaBERT (Marathi), BanglaBERT (Bengali), etc. It is noticed that the monolingual MahaBERT model performs the best on HASOC-Marathi, whereas the multilingual MuRILBERT performs the best on HASOC-Hindi and BenHateSpeech. Several other cross-language evaluations over Marathi and Hindi monolingual models and mixed observations are presented.

1 Introduction

According to the Cambridge Dictionary, hate speech is defined as - “Hate speech is a public speech that expresses hate or encourages violence towards a person or group based on race, religion, sex, or sexual orientation”[1]. Statistics reveal that half of the world’s population, including print media, is now engaged in social media platforms[2] and 12½ trillion hours spent online by the users[3]. This trend shall continue till obvious infinity. Sometimes aggressive posts, misleading news, and harassing comments can lead people to social violence, even riots (Laub, 2019). Worldwide, Governments are introducing laws against hate speech. So, digital media like Twitter, Facebook, etc., are also becoming more concerned about it and endeavouring to filter hate, sexual abuse, harmful acts, harassment, bullying, child abuse, etc.

Researchers explore this field, but most of the experiment is based on European language datasets[4]. Limited work is done on the Indian languages except for publishing datasets or improving accuracy. India has 22 official languages and about 1,000 living languages from various language groups (Kalra and Dutt, 2019). People in India use their native lan-

[1] https://dictionary.cambridge.org/dictionary/english/hate-speech
[2] https://datareportal.com/reports/digital-2021-global-overview-report
[3] https://datareportal.com/reports/digital-2022-global-overview-report
[4] https://hatespeechdata.com/

guages on social media platforms, and sometimes users don't follow the proper structure or grammar, making it more complicated to detect hate speech in the computational aspect. This situation motivated us to work on hate speech on Twitter and other social media texts. It is challenging for automatic approaches to detect hate speech in text.

Researchers use state-of-the-art transformer models more in language-related researchs like NLP, Information Retrieval (IR), etc., to enrich performance. Many works have already been done in NLP like text classification (Sun et al., 2019), question-answering (McCarley et al., 2019), token classification (Ulčar and Robnik-Šikonja, 2020), and Named Entity Recognition (NER) (Luoma and Pyysalo, 2020). The pre-trained BERT-based masked language models have been used, and these language models' multilingual and monolingual variants have drawn attention to the low-resource languages.

This paper attempted to identify hate speech content in Hindi, Marathi and Bengali comments collected from social media. We choose various publicly available datasets like HASOC (Hate Speech and Offensive Content Identification)[5] and Bangla Hate Speech datasets (BenHateSpeech)[6] with binary classification. Variation of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), RoBERTa (Robustly optimized BERT) (Liu et al., 2019), ALBERT (A Lite BERT) (Lan et al., 2019), DistilBERT (Distilled version of BERT) (Sanh et al., 2019) and their pre-trained models such as mBERT (Devlin et al., 2018), MuRIL-BERT (Khanuja et al., 2021), NeuralSpace-BERTHi (Jain et al., 2020), RoBERTa-Hindi, Indic Bert (Kakwani et al., 2020), MahaBERT (Joshi, 2022a), MahaRoBERTa (Joshi, 2022b), XLM-RoBERTa (Conneau et al., 2019), BanglaBert (Sarker, 2020) etc. is used for this work.

Later, we compare different pre-trained BERT architecture's performance on publicly available datasets in Hindi, Marathi and Bengali languages. We also compare multilingual and monolingual variants of these language models. The monolingual models used here are only pre-trained on Hindi, Bengali, or Marathi data. Next, we follow a cross-language evaluation of these BERT models (Litake et al., 2022) since both Hindi and Marathi share the Devanagri script.

Our focus in this work is :

- To see how well pre-trained BERT models perform, utilize the mono and multilingual pre-trained BERT models with their variants.
- A detailed comparison between all the models for Hindi, Marathi and Bengali languages has been made. Almost thirty experiments have been done, twelve for Hindi, Marathi and six for Bengali.
- There is a cross-language experiment between Hindi and Marathi where monolingual Marathi models, i.e. MahaBert, MahaRoBERTa, RoBERTa-Base-Mr, and MahaAlBERT performs well on the Hindi dataset. NeuralspaceBERTHi, Roberta-Hindi and DistilBERTHindi, which are monolingual Hindi models, also perform well in the case of the Marathi dataset.

The rest of the paper is structured as follows. Section 2 is the work related to hate speech detection in Indian languages. Section 3 describes the experimental setup like the dataset, preprocessing steps, BERT variants and pre-trained models. Section 4 summarizes the results and findings from all of the experiments. Finally, it is concluded in Section 5.

[5] https://hasocfire.github.io/hasoc/2022/index.htm

[6] https://www.kaggle.com/naurosromim/bengali-hate-speech-dataset

2 Related Work

There is very little research on the Indian hate speech dataset as less data is available publicly. Creating labelled datasets of hate speech in the Indian language is tedious and challenging. It needs lots of groundwork and preprocessing, like cleaning, annotators' agreements, etc., to create valuable data from social media. In this section, language-wise, we shall discuss some existing datasets and the work done on those datasets.

Hindi: HASOC (Hate Speech and Offensive Content Identification), a shared task organized by FIRE (Forum for Information Retrieval Evaluation)[7], which published hate datasets in Indian languages such as Hindi, Marathi, etc. HASOC offers four subtracks, one of which is relevant to us: **HASOC - English and Indo-Aryan Languages**. Datasets are distributed in tab-separated format. HASOC and most other collections require mechanisms to detect hateful content from the text of a post.

In 2019, the HASOC-Hindi dataset offered three tasks (Mandl et al., 2019). The first task is binary classification, i.e. subtask A. The second task is to find whether the hate comment was profane or abusive (multiclass), i.e. subtask B. The third is to predict whether the hate comment is targeted or untargeted (multiclass), i.e. subtask C. In the Hindi language, ninety-three runs were submitted across three subtasks. Regarding the Hindi subtask A, the winner team, QutNocturnal (Bashar and Nayak, 2020), employed a CNN base technique with Word2vec embedding and got better Marco F1 and Weighted F1 values, 0.8149 and 0.8202, respectively. The second team LGI2P (Mensonides et al., 2019), trained a fastText model for the proposed Hindi language and later used BERT for classification. The system achieved 0.8111 Marco-F1 and 0.8116 Weighted-F1 values. For sub-task B on Hindi Dataset, 3Idiots (Mishra and Mishra, 2019) scores 0.5812 and 0.7147 in Marco-F1 and Weighted-F1 utilizing BERT. Team A3-108 (Mujadia et al., 2019) achieves a high Marco-F1 score on subtask C Hindi Dataset, which is 0.5754. According to them, Adaboost (Freund and Schapire, 1997) was the best performing classifier among the three classifiers, i.e., Adaboost or Adaptive Boosting (AB), Random Forest (RF), Linear Support Vector Machine (SVM). They merge multiple weak classifiers to construct a robust prediction model, but an ensemble of SVM, Random Forest, and Adaboost with hard voting performed even better. This classifier used TF-IDF features of word unigrams and characters 2, 3, 4, and 5 grams with an additional feature of the length of every tweet.

In HASOC 2020, two Hate Speech detection tasks (Mandl et al., 2020), sub-task A (binary class) and sub-task B (multiclass), are proposed with another Hindi dataset in the research area. NSIT_ML_Geeks (Raj et al., 2020) outperforms other teams in the competition scoring Marco-F1 0.5337 and 0.2667 in sub-task A and sub-task B, respectively utilizing CNN and BiLSTM. Nohate (Kumari) team achieved Marco-F1 0.3345 in sub-task B, fine-tuning BERT model for the classification.

In 2021, HASOC published a Hindi dataset (Modha et al., 2021) with sub-task A and B again. Total Sixty-five teams submitted a total of six thousand and fifty-two runs. The best submission was achieved Macro F1 0.7825 in sub-task A with a fine-tuned Multilingual-BERT (20 epochs) with a classifier layer added at the final phase. The second team also fine-tuned Multilingual-BERT and scored Macro F1 0.7797. NeuralSpace (Bhatia et al., 2021) got Macro F1 0.5603 in sub-task B. They use an XLM-R transformer, vector representations for emojis using the system Emoji2Vec, and sentence embeddings for hash-

[7] http://fire.irsi.res.in/fire/2022/home

tags. After that, three resulting representations were concatenated before classification.

In the paper (Bhardwaj et al., 2020) they used the pre-trained multilingual BERT (m-BERT) model for computing the input embedding on the Hostility Detection Dataset (Hindi) later SVM, Random-Forest, Multilayer perceptron (MLP), Logistic Regression models are used as classifiers. In coarse-grained evaluation, SVM reported the best weighted-F1 score of 84%, whereas they obtained 84%, 83%, and 80% weighted-F1 scores for LR, MLP, and RF. In fine-grained evaluation, SVM has the most excellent F1 score for evaluating three hostile dimensions, namely Hate (47%), Offensive (42%), and Defamation (43%). Logistic Regression beats the others in the Fake dimension with an F1 score of 68%.

Marathi: In HASOC-Marathi (Modha et al., 2021), the best-performing team WLV-RIT fine-tuned XLM-R Large model with a simple softmax layer. Later executed transfer learning from English data released for OffensEval 2019 (Zampieri et al., 2019) and Hindi data released for HASOC 2019 (Mandl et al., 2019) and show that executing transfer learning from Hindi is better than executing transfer learning from English. They Scored an F1 score of 0.9144 (Nene et al., 2021). The second team applied a fine-tuned LaBSE transformer (Feng et al., 2020) on the Marathi data set and the Hindi data set and achieved an F1 score of 0.8808. Their experiments show that the LaBSE transformer (Glazkova et al., 2021) outperforms XLM-R in the monolingual settings, but XLM-R performs better when Hindi and Marathi data are merged. L3CubeMahaHate (Velankar et al., 2022) presents the first major Marathi hate speech dataset with 25,000 distinct tweets from Twitter, later annotated manually, and labelled them into four major classes, i.e. hate, offensive, profane, and not. Finally, they use CNN, LSTM, and Transformers. Next, they explore monolingual and multilingual variants of BERT like MahaBERT, IndicBERT, mBERT, and xlm-RoBERTa and show that monolingual models perform better than their multilingual counterparts. Their MahaBERT (Joshi, 2022a) model provides the best results on L3Cube-MahaHate Corpus.

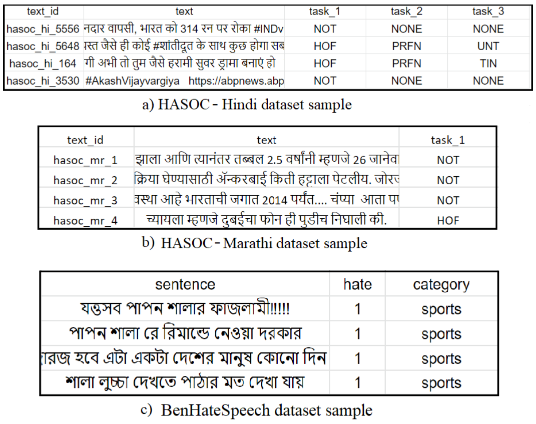
Bengali: Karim et al. (Karim et al., 2020) published a Bengali dataset with 35,000 hate statements (political, personal, geopolitical, and religious) and applied a multichannel CNN and LSTM-based approach. Later DeepHate-Explainer (Karim et al., 2021) added more than 5,000 labelled examples with it and used an ensemble method of transformer-based neural architectures to classify them into political, personal, geopolitical, and religious hates and achieved F1-scores of 78%, 91%, 89%, and 84%, for political, personal, geopolitical, and religious hates. In the paper (Romim et al., 2021), They published a Bengali Hate Speech corpus with 30,000 comments labelled with “1” for hate comments; otherwise, “0”. This paper (Mandal et al., 2022) created a political news corpus and then developed a keyword or phrase-based hate-speech identifier using a semi-automated approach.

Most of the top results are delivered by the systems based on Deep neural models and transformers.

3 Experimental Setup

3.1 Dataset Selection

The experiment uses the HASOC-Hindi (2019), HASOC-Marathi (2021) and BenHate-Speech (Romim et al., 2021) datasets. Statistics and class distribution for the training set of HASOC-Hindi, HASOC-Marathi and BenHateSpeech datasets are in table 1 and for the test set in table 2. Figure 1 shows samples of datasets, and we chose only binary classification task for our work, i.e., to detect whether a sentence or text conveys hate or not. For HASOC-Hindi and HASOC-Marathi, classes are "HOF" and "NOT" whereas in BenHateSpeech, classes are "1" (hate) and "0" (not).



text_id	text	task_1	task_2	task_3
hasoc_hi_5556	नदार वापसी, भारत को 314 रन पर रोका #INDv	NOT	NONE	NONE
hasoc_hi_5648	स्त जैसे ही कोई #शांतीदूत के साथ कुछ होगा सब	HOF	PRFN	UNT
hasoc_hi_164	गी अभी तो तुम जैसे हरामी सुवर ड्रामा बनाएं हो	HOF	PRFN	TIN
hasoc_hi_3530	#AkashVijayvargiya https://abpnews.abp	NOT	NONE	NONE

a) HASOC - Hindi dataset sample

text_id	text	task_1
hasoc_mr_1	झाला आणि त्यानंतर तब्बल 2.5 वर्षांनी म्हणजे 26 जानेवा	NOT
hasoc_mr_2	क्रिया घेण्यासाठी अँन्करबाई किती हट्टाला पेटलीय. जोरज	NOT
hasoc_mr_3	वस्था आहे भारताची जगात 2014 पर्यंत.... चंप्या आता प	NOT
hasoc_mr_4	च्यायला म्हणजे दुबईचा फोन ही पुडीच निघाली की.	HOF

b) HASOC - Marathi dataset sample

sentence	hate	category
যত্তসব পাপন শালার ফাজলামী!!!!!	1	sports
পাপন শালা রে রিমান্ডে নেওয়া দরকার	1	sports
ারজ হবে এটা একটা দেশের মানুষ কোনো দিন	1	sports
শালা লুচ্চা দেখতে পাঠার মত দেখা যায়	1	sports

c) BenHateSpeech dataset sample

Figure 1: Samples of HASOC-Hindi, HASOC-Marathi and BenHateSpeech datasets respectively

3.2 Preprocessing

Before employing data to the transformers, text data need to be cleaned and noise-free to enrich the performance. India's low-resourced languages share several characteristics. Despite being written in different languages, researchers employed nearly identical preprocessing approaches for all datasets. Text preprocessing procedures can be slightly different depending on the task and the dataset used. Some datasets had raw comments with emojis, punctuation, and unwanted characters. In most cases, the following steps are used:

Normalization: It defines the removal of existing emojis, unwanted characters and stopwords from the sentences.

Removing punctuation and number: Punctuation and numbers often don't add extra meaning to the text hence being removed from the text.

Word tokenization : Here converting a sentence into an individual word is called a token.

Stemming: Stemming removes the inflections from each word to convert that word to its root word.

Label encoding: In HASOC-Hindi and HASOC-Marathi, task_1 is tagged as "NOT" and "HOF". We encode them into a unique number. Like, "NOT" to "0" and "HOF" to "1", where we leave BenHateSpeech dataset as it is.

In the case of HASOC-Marathi and BenHateSpeech datasets, we followed above mentioned simple steps. In case of HASOC-hindi, we followed preprocessing techniques as mentioned in paper (Bashar and Nayak, 2020) like replacing person occurrence (e.g. @someone) with `xxatp`, URL occurrence with `xxurl`, source of modified retweet with `xxrtm`, source of not modified retweet with `xxrtu`, fixing the repeating characters (e.g. goooood), removed common invalid characters (e.g. $< br =>$, $< unk >$, $@ - @$,etc) and a lightweight stemmer for Hindi language (Ramanathan and Rao, 2003) for stemming the words.

3.3 Transformer Language Models

Figure 2 shows a general transformer-based BERT model structure, and the input text is in the Bengali language. After the above steps of preprocessing, we employ the p number of texts of the training set (D) is indicated as

$$D = \{T_1, T_2, T_3, .., T_i, ...T_p\},$$

where T_i is the i^{th} number of texts and p is equal to the total number of texts present in a training set. Given a text T_i, the text having m words, i.e., length of the text, is denoted as

$$T_i = \{w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,k}, .., w_{i,m}\},$$

where $w_{i,k}$ denotes the k^{th} word in the i^{th} text.

Datasets	HOF/Hate	NOT	Total
HASOC-Hindi (2019)	2,469	2,196	4,665
HASOC-Marathi (2021)	669	1,205	1,874
BenHateSpeech	8,000	16,000	24,000

Table 1: Class distribution analysis for training set

Datasets	HOF/Hate	NOT	Total
HASOC-Hindi (2019)	605	713	1318
HASOC-Marathi (2021)	207	418	625
BenHateSpeech	2,000	4,000	6,000

Table 2: Class distribution analysis for test set

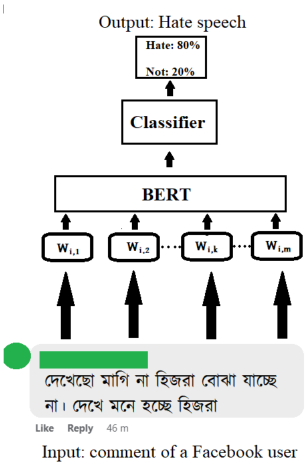


Figure 2: A general transformer-based BERT model architecture

3.3.1 BERT

BERT is developed by Google, a transformer-based technique for NLP. BERT can generate contextualized embeddings. It produces almost similar vectors for synonyms and different vectors if the use of words is different. During training, it learns the details from both sides of the word's context. So, it is called a bidirectional model. We evaluated mono and multilingual BERT on Hindi, Marathi and Bengali datasets. Due to memory and *GPU* issues, we did several experiments but with the same hyperparameter combination (Table 3).

mBERT[8]**:** It is pre-trained with the largest Wikipedia over 104 top languages worldwide, including Hindi, Bengali and Marathi, using a masked language modelling (MLM) objective.

MuRILBERT[9]**:** Multilingual Representations for Indian Languages (MuRIL) is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts, i.e. monolingual segments and parallel segments.

NeuralspaceBERTHi[10]**:** This BERT model is pre-trained on approx. 3 GB of monolingual training corpus, i.e., OSCAR corpus released by neuralspace-reverie. It fine-tuned downstream tasks like text classification, POS-tagging, question-answering, etc.

MahaBERT[11]**:** MahaBERT is a multilingual BERT (bert-base-multilingual-cased) model finetuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets.

BanglaBERT[12]**:** Using mask language modelling, bangla-Bert-Base was pre-trained on data downloaded from OSCAR and Bengali Wikipedia Dump Dataset.

[8] https://huggingface.co/bert-base-multilingual-cased

[9] https://huggingface.co/google/muril-base-cased

[10] https://huggingface.co/neuralspace-reverie/indic-transformers-hi-bert

[11] https://huggingface.co/l3cube-pune/marathi-bert

[12] https://huggingface.co/sagorsarker/bangla-bert-base

Hyperparameter	BERT variants
Learning-rate	$1e$-5
Epochs	5
Max seq length	512
Batch size	8

Table 3: Combination of hyperparameters for training BERT variants

3.3.2 RoBERTa

More extended time training on a large dataset can increase BERT's performance. This model, called RoBERTa, a self-supervised transformer model trained on raw texts, outperforms BERT by 4%-5% on natural language inference and utilizes a character-level BPE (Byte Pair Encoding) tokenizer. Still, RoBERTa uses a byte-level BPE tokenizer, which benefits from a universal encoding scheme.

XLM-RoBERTa (base-sized model)[13]: The XLM-RoBERTa model is a multilingual version of the RoBERTa model pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages. It does not require $lang$ tensors like XLM multilingual models to determine which language is utilized and choose the correct language based on the input ids.

Roberta-Hindi[14]: This RoBERTa transformer base model was pre-trained on a large Hindi corpus (a combination of MC4, OSCAR, and indic-nlp datasets) released by flax-community.

MahaRoBERTa[15]: A Multilingual RoBERTa (xlm-roberta-base) model fine-tuned on publicly available Marathi monolingual datasets and L3Cube-MahaCorpus.

RoBERTa-Base-Mr[16]: The RoBERTa Marathi model was pre-trained on *mr* dataset of C4 (Colossal Clean Crawled Corpus) (Raffel et al., 2019) multilingual dataset.

3.3.3 ALBERT:

A lite BERT for self-supervised learning, Google AI open-sourced ALBERT uses fewer parameters than BERT.

IndicBERT[17]: IndicBERT trained on large-scale datasets is a multilingual ALBERT model covering 12 major Indian languages (such as Hindi, Marathi, Bengali, Assamese, English, Gujarati, Oriya, Punjabi, Tamil, Telugu, Kannada and Malayalam) released by Ai4Bharat.

MahaALBERT[18]:A Marathi ALBERT model trained on publicly available Marathi monolingual datasets and L3Cube-MahaCorpus.

3.3.4 DistilBERT:

DistilBERT is a small, quick, inexpensive, and light transformer model trained by distilling the BERT base. More than 95% of BERT's performance on the GLUE language understanding benchmark is preserved in this version, which has 40% fewer parameters and runs 60% faster.

mDistilBERT[19]: The model is trained on the concatenation of 104 different languages of Wikipedia.

DistilBERTHindi[20]: A DistilBERT language model pre-trained on approx. 10 GB of monolingual training corpus, which is taken from OSCAR.

[13] https://huggingface.co/xlm-roberta-base

[14] https://huggingface.co/flax-community/roberta-hindi

[15] https://huggingface.co/l3cube-pune/marathi-roberta

[16] https://huggingface.co/flax-community/roberta-base-mr

[17] https://huggingface.co/ai4bharat/indic-bert

[18] https://huggingface.co/l3cube-pune/marathi-albert-v2

[19] https://huggingface.co/distilbert-base-multilingual-cased

[20] https://huggingface.co/neuralspace-reverie/indic-transformers-hi-distilbert

4 Result and Analysis

In this section, we discuss the precision, recall and weighted F1 score obtained by training all the variants of BERT on Hindi, Marathi and Bengali datasets. Table 4 represents the results of transformer models trained on the HASOC-Hindi, HASOC-Marathi, and Ben-HateSpeech datasets, where blue, purple, and teal colours indicate multilingual, monolingual, and cross-language models correspondingly. We intentionally prefer a weighted F1 score over an accuracy score to evaluate the models because imbalanced class distribution exists in most classification problems. So, weighted F1 score is a better metric to consider in this scenario. In the training set, the number of sentences for Bengali is double that of Hindi and Marathi. The models trained on the Bengali dataset have better results than Marathi and Hindi. To evaluate our models, we use two class precisions (P_{NOT}, P_{HOF}), recalls (R_{NOT}, R_{HOF}), F1 scores ($F1_{NOT}$, $F1_{HOF}$) then calculate weighted precision (W_P), recall (W_R), and F1 score (W_{F1}) here. At last we calculate $Accuracy$.

$$P_{NOT} = \frac{True_{NOT}}{True_{NOT} + False_{HOF}} \tag{1}$$

$$P_{HOF} = \frac{True_{HOF}}{True_{HOF} + False_{HOF}} \tag{2}$$

$$R_{NOT} = \frac{True_{NOT}}{True_{NOT} + False_{NOT}} \tag{3}$$

$$R_{HOF} = \frac{True_{HOF}}{True_{HOF} + False_{NOT}} \tag{4}$$

$$F1_{NOT} = 2 * \frac{P_{NOT} * R_{NOT}}{P_{NOT} + R_{NOT}} \tag{5}$$

$$F1_{HOF} = 2 * \frac{P_{HOF} * R_{HOF}}{P_{HOF} + R_{HOF}} \tag{6}$$

$$W_P = \frac{P_{NOT} * T_{NOT} + P_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{7}$$

$$W_R = \frac{R_{NOT} * T_{NOT} + R_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{8}$$

$$W_{F1} = \frac{F1_{NOT} * T_{NOT} + F1_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{9}$$

$$Accuracy = \frac{True_{NOT} + True_{HOF}}{T_{NOT} + T_{HOF}} \tag{10}$$

Where $True_{NOT}$ = True-negative (model predicted the texts as NOT, and the actual value of the same is also NOT), $True_{HOF}$ = True-positive (model predicted the texts as HOF, and the actual value of the same is also HOF), $False_{NOT}$ = False-negative (model predicted the texts as NOT, but the true value of the same is HOF), $False_{HOF}$ = False-positive (model predicted the texts as HOF, but the true value of the same is NOT), P_{NOT} = Precision of NOT class, P_{HOF} = Precision of HOF class, R_{NOT} = Recall of NOT class, R_{HOF} = Recall of HOF class, $F1_{NOT}$ = F1 score of NOT class, $F1_{HOF}$ = F1 score of HOF class, T_{NOT} = The total number of NOT class text present in test set, T_{HOF} = The total number of HOF class text present in test set

Best models per datasets: The weighted F1 score for the top four models like MuRIL-BERT, MahaRoBERTa, NeuralSpaceBERTHi, and XLM-RoBERTa are very close for the Hindi dataset. MahaBERT, MahaRoBERTa, mBERT, and Roberta-Hindi score top for the Marathi dataset. MuRILBERT, BanglaBert, and XLM-RoBERTa models are most suitable for the Bengali dataset. Figure 3 shows the confusion matrix of the best models on three datasets separately.

Monolingual models vs multilingual models: On the Hindi dataset, multilingual models like MuRILBERT and XLM-RoBERTa perform better, but the monolingual model NeuralSpaceBERTHi also gives tough competition. We can conclude that multilingual models perform well, but the difference in performance between monolingual and multilingual models is negligible. MahaBERT and MahaRoBERTa models provide the highest weighted F1 score

Models on HASOC (Hindi)	**Precision**			**Recall**			**F1 score**			**Accuracy**
	0	**1**	**w.avg.**	**0**	**1**	**w.avg.**	**0**	**1**	**w.avg.**	
mBERT	0.8078	0.7797	0.7949	0.8275	0.8016	0.8156	0.8175	0.7904	0.8050	0.8050
MuRILBERT	0.8695	0.8362	0.8542	0.8266	0.7851	0.8075	0.8475	0.8098	**0.8301**	**0.8308**
NeuralSpaceBERTHi	0.8611	0.8278	0.8458	0.8263	0.7867	0.8081	0.8433	0.8067	0.8264	0.8270
MahaBERT	0.8681	0.8297	0.8504	0.8080	0.7570	0.7845	0.8369	0.7916	0.8161	0.8171
XLM-RoBERTa	0.8218	0.7977	0.8107	**0.8492**	**0.8280**	**0.8394**	0.8352	**0.8125**	0.8247	0.8247
Roberta-Hindi	0.8485	0.8147	0.8329	0.8231	0.7851	0.8056	0.8356	0.7996	0.8190	0.8194
MahaRoBERTa	**0.8892**	**0.8534**	**0.8727**	0.8138	0.7603	0.7892	**0.8498**	0.8041	0.8288	0.8300
RoBERTa-Base-Mr	0.8246	0.7906	0.8089	0.8155	0.7801	0.7992	0.8200	0.7853	0.8040	0.8042
IndicBERT	0.7489	0.7198	0.7355	0.7864	0.7603	0.7744	0.7671	0.7394	0.7543	0.7541
MahaAlBERT	0.8232	0.7913	0.8085	0.8221	0.7900	0.8073	0.8226	0.7906	0.8079	0.8080
mDistilBERT	0.7812	0.7487	0.7662	0.7991	0.7685	0.7800	0.7900	0.7584	0.7754	0.7754
DistilBERTHindi	0.8064	0.7781	0.7934	0.8261	0.8000	0.8141	0.8161	0.7888	0.8035	0.8034
Models on HASOC (Marathi)										
mBERT	0.9019	0.8110	0.8717	**0.9240**	**0.8502**	**0.8995**	0.9128	0.8301	0.8854	0.8848
MuRILBERT	0.8995	0.7878	0.8625	0.8805	0.7536	0.8384	0.8898	0.7703	0.8502	0.8512
NeuralSpaceBERTHi	0.9066	0.8115	0.8751	0.9066	0.8115	0.8751	0.9066	0.8115	0.8751	0.8752
MahaBERT	0.9234	0.8415	0.8962	0.9125	0.8212	0.8822	**0.9179**	**0.8312**	**0.8891**	**0.8896**
XLM-RoBERTa	0.8588	0.7242	0.8142	0.8734	0.7487	0.8320	0.8660	0.7336	0.8221	0.8224
Roberta-Hindi	0.9354	0.8540	0.9084	0.8886	0.7632	0.8470	0.9113	0.8060	0.8764	0.8784
MahaRoBERTa	0.9306	0.8520	0.9045	0.9067	0.8067	0.8735	0.9184	0.8287	0.8886	**0.8896**
RoBERTa-Base-Mr	**0.9688**	**0.8960**	**0.9446**	0.8100	0.5410	0.7209	0.8823	0.6746	0.8135	0.8272
IndicBERT	0.8708	0.6785	0.8071	0.7964	0.5507	0.7150	0.8319	0.6079	0.7577	0.7648
MahaAlBERT	0.9138	0.8095	0.8792	0.8761	0.7391	0.8307	0.8945	0.7726	0.8541	0.8560
mDistilBERT	0.8588	0.6878	0.8021	0.8233	0.6280	0.7586	0.8406	0.6565	0.7796	0.7824
DistilBERTHindi	0.9066	0.7989	0.8709	0.8793	0.7487	0.8360	0.8927	0.7729	0.8530	0.8544
Models on BenHateSpeech										
mBERT	0.9303	0.8630	0.9078	0.9155	0.8362	0.8890	0.9228	0.8493	0.8983	0.8980
MuRILBERT	0.9225	0.8507	0.8985	**0.9406**	**0.8835**	**0.9215**	**0.9314**	**0.8667**	**0.9098**	**0.9095**
XLM-RoBERTa	**0.9463**	**0.8975**	**0.9300**	0.9047	0.8249	0.8781	0.9250	0.8596	0.9032	0.9023
IndicBERT	0.9042	0.8030	0.8704	0.9300	0.8515	0.9038	0.9169	0.8265	0.8867	0.8876
BanglaBERT	0.9333	0.8685	0.9117	0.9207	0.8456	0.8956	0.9269	0.8568	0.9035	0.9033
mDistilBERT	0.8698	0.7290	0.8228	0.9055	0.7941	0.8683	0.8872	0.7601	0.8448	0.8466

Table 4: Precision, Recall, F1 score, and Accuracy of various transformer models on HASOC-Hindi, HASOC-Marathi and BenHateSpeech datasets, respectively

for the Marathi dataset, and mBERT also performs well, whereas MuRILBERT scores a little less. For Bengali, we use only one monolingual pre-trained model, i.e., BanglaBERT; it performs very well, but the MuRILBERT wins marginally. IndicBERT and mDistilBERT models' performance is significantly less on all datasets than in other models. Therefore, developing better resources for the Hindi and Bengali language is necessary as language-specific fine-tuning does not necessarily guarantee the best performance.

Cross-language experiments: During the cross-language experiments, we consider the Marathi models on the Hindi dataset and vice-versa, as both languages share the Devanagari script. MahaRoBERTa performs pretty well on the Hindi dataset, and MahaBERT, RoBERTa-Base-Mr, and MahaALBERT also score sufficiently. NeuralSpaceBERTHi and Roberta-Hindi perform well on the Marathi dataset, but surprisingly, DistilBERTHindi performs poorly on the Hindi dataset rather well on the Marathi dataset.

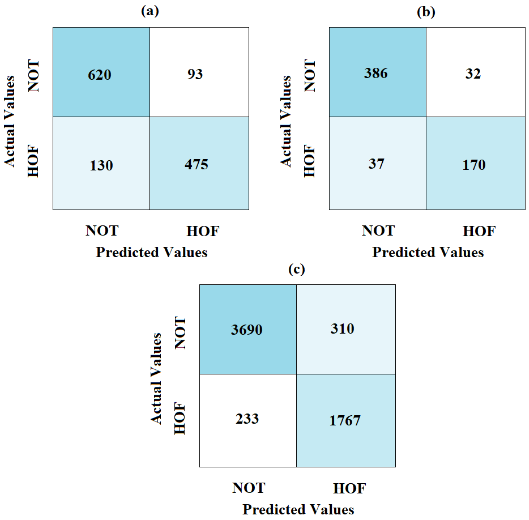


Figure 3: Confusion matrix of Best models such as MuRILBERT for Hindi (a), MahaBERT for Marathi (b) and MuRILBERT for Bengali (c)

5 Conclusion and Future Scope

Significant works are done in English or other major speaking languages. In Indian languages, a very little work has been done, like on Hindi, Bengali, Marathi, Tamil, Malayalam etc. In brief, we conclude this work by: (i) Exploring variants of transformer-based models in Indic languages. (ii) Comparing monolingual and multilingual transformer-based models for hate speech detection data like HASOC-Hindi, HASOC-Marathi and BenHateSpeech. (iii) Cross-language experiments on Hindi and Marathi models. Results show that monolingual training doesn't necessarily ensure superior performance. Multilingual models stood first on Bengali and Hindi datasets, whereas Marathi monolingual models performed the best on Marathi dataset. Our next concentration will be to reduce false-positive and false-negative errors. We also observe that the "0" class precision, recall, and F1 score is slightly higher than the "1" class, indicating the data imbalance. So, in future, techniques like SMOTE (Bowyer et al., 2011), ADASYN (He et al., 2008), or data augmentation (techniques to increase the amount of data) (Nozza, 2022) can be used, which can handle data imbalance. Apart from the technical challenges, the research on hate speech impacts other dimensions, including socio-linguistic issues like freedom of speech and legislation at the national and international levels. Socio-linguistic implications of this research are that some word is used to target a few specific castes, community, colour, etc. A sophisticated hate speech detection system can identify such hated information, restrict its propagation, and alert concerned authorities. In a social context, freedom of speech is related to this issue. So there must be some trade-off between them to maintain peace and harmony.

References

Md. Abul Bashar and Richi Nayak. 2020. Qutnocturnal@hasoc'19: CNN for hate speech and offensive content identification in hindi language. *CoRR*, abs/2008.12448.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *arXiv preprint arXiv:2011.03588*.

Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Prakash Ramesh, Shubham Gupta, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2021. One to rule them all: Towards joint indic language hate speech detection. *CoRR*, abs/2109.13711.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Anna Glazkova, Michael Kadantsev, and Maksim Glazkov. 2021. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. *arXiv preprint arXiv:2110.12687*.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi BERT language models, and resources. *CoRR*, abs/2202.01159.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of The WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101, Marseille, France. European Language Resources Association.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Rajrani Kalra and Ashok K. Dutt. 2019. Exploring linguistic diversity in india: A spatial analysis. *Handbook of the Changing World Language Map*.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

S. Kumari. Nohate at hasoc2020: Multilingual hate speech detection. In *Forum for Information Retrieval Evaluation*, FIRE 2020.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Zachary Laub. 2019. Hate speech on social media: Global comparisons. *Council on foreign relations*, 7.

Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. 2022. Mono vs multilingual bert: A case study in hindi and marathi named entity recognition.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with bert. *arXiv preprint arXiv:2006.01563*.

Prasanta Mandal, Apurbalal Senapati, and Amitava Nag. 2022. Hate-speech detection in news articles: In the context of west bengal assembly election 2021. In *Pattern Recognition and Data Analysis with Applications*, pages 247–256, Singapore. Springer Nature Singapore.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

JS McCarley, Rishav Chakravarti, and Avirup Sil. 2019. Structured pruning of a bert-based question answering model. *arXiv preprint arXiv:1910.06360*.

Jean-Christophe Mensonides, Pierre-Antoine Jean, Andon Tchechmedjiev, and Sébastien Harispe. 2019. Imt mines ales at hasoc 2019: automatic hate speech detection. In *FIRE 2019-11th Forum for Information Retrieval Evaluation*, volume 2517, pages p–279.

Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In *FIRE (Working Notes)*, pages 208–213.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 1–3, New York, NY, USA. Association for Computing Machinery.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2019. Iiit-hyderabad at hasoc 2019: Hate speech detection. In *FIRE (Working Notes)*, pages 271–278.

Mayuresh Nene, Kai North, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Transformer models for offensive language identification in marathi. In *FIRE*.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Roushan Raj, Shivangi Srivastava, and Sunil Saumya. 2020. Nsit & iiitdwd @ hasoc 2020: Deep learning model for hate-speech identification in indo-european languages. In *FIRE*.

Ananthakrishnan Ramanathan and Durgesh Rao. 2003. A lightweight stemmer for hindi.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, Saiful Islam, et al. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert. In *International Conference on Text, Speech, and Dialogue*, pages 104–111. Springer.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and BERT models. *CoRR*, abs/2203.13778.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.